Towards understanding how to instruct students in dichotomous identification keys in a mixed STEM forensic science education environment.

Trevor Stamper[1]* Ph.D., Lauren M. Weidner[1†] Ph.D., Gregory Nigoghosian[1]M.S., Nastasha Johnson[2], Cong Wang[3††] Ph.D., and Chantal Levesque-Bristol[3] Ph.D.

*[1] Department of Entomology, Purdue University, West Lafayette, IN 47907, USA; *corresponding author: stampert@purdue.edu*
[2] University Libraries and School of Information Science, Purdue University, West Lafayette, IN 47907, USA
[3]Center for Instructional Excellence, Purdue University, West Lafayette, IN 47907, USA
†Present Address: School of Mathematical and Natural Sciences, Arizona State University, 4701 W Thunderbird Rd, Glendale, AZ 85306, USA
††Present Address: Department of Ecology and Evolutionary Biology, Yale University, OML 301, 165 Prospect St., New Haven, CT 06511

Abstract: Morphological assessment is a traditional approach to specimen identification in many forensic subdisciplines. A dichotomous key guides the user through taxa determination for a specimen by providing a series of choice nodes that center around morphological differences. Each nodal choice leads to either a new set of dichotomous choices or a taxa decision. In a forensic analysis course, we evaluated student's ability to utilize a dichotomous key down to species for a limited set of taxa, by reviewing their nodal decisions along with their confidence level using a Likert scale (1-5). Along with individual decision recording, students conducted a post-decision group comparison, following a think-pair-share active learning model. If student answers were not the same, they re-evaluated their specimen until a mutual evidence-based decision was reached. Students displayed high decision confidence but low accuracy. We observed a higher initial accuracy from students enrolled in STEM majors when compared to non-STEM majors. From these data we aim to improve student training in the use of dichotomous keys for species identification, with a continued approach that can be then used to provide guidelines for how forensic scientists should approach dichotomous key training.

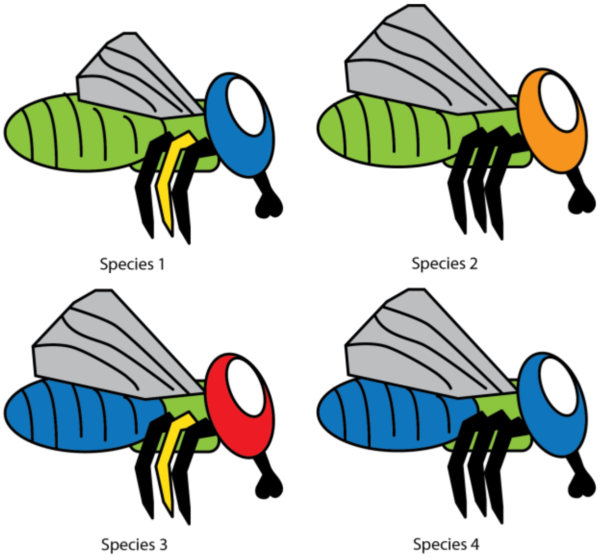
Keywords: Insect Identification, Entomology, Classification, Key character

.

Introduction

In science, the ability to quickly identify a specimen with accuracy and precision is a challenge. Identification keys are a central cataloging and naming tool for diverse groups of organisms, such as: animals (e.g.—1), plants (e.g.—2), and even pollen grains (e.g.—3). Identification keys hold a lot in common with decision trees, guiding the individual to a final decision based on criteria and decisions. Such keys are not limited to just extant species data, having been used to connect fossils with living groups (4). Further, identification keys are applicable to even non-biological groups, such as soil types (5, 6), minerals (7, http://www.minsocam.org/msa/collectors_corner/id/mineral_id_keyi1.htm), and anthropological artifacts (e.g.—http://www.projectilepoints.net).

Forensically, identification keys are utilized in many ways, such as: fingerprints (8), skeletal osteology (9), entomological evidence (10), and even presumptive drug testing (11). Exposure and training in using identification keys in forensic science is an important curricular consideration when teaching students scientific analysis. This is especially true considering forensic science straddles STEM and Non-STEM as it strives to bring together academic and practitioner viewpoints. Furthermore, the use of identification keys is something that is best learned by doing, since it involves direct observation skills that require practice to perfect.



single-access key

Node	Couplet Description	Decision
1a	Abdomen color green	2
1b	Abdomen color blue	3
2a	Middle leg yellow; small wings	Sp. 1
2b	Middle leg black; big wings	Sp. 2
3a	Head color red	Sp. 3
3b	Head color blue	Sp. 4

Figure 1a

multi-access key

Species	Abdomen Color	Middle Leg Color	Wing Size	Head Color
Sp. 1	Green	Yellow	Small	Blue
Sp. 2	Green	Black	Big	Orange
Sp. 3	Blue	Yellow	Big	Red
Sp. 4	Blue	Black	Big	Blue

Figure 1b

FIGURE 1: *Example single-access (1a) and multi-access (1b) key for four fly species. Each species displays different character states, denoted by either size or color differences in the schematics above (species 1-4; species is abbreviated as sp.). Both keys allow for species-level diagnosis, but do so in different ways. The single-access key provides a structured set of pre-constructed decisions (denoted in couplets: 1a+1b, 2a+2b, 3a+3b) that guide the user through the identification. The multi-access key provides the same data, but in an unconstructed format, with no guidance through the process of identification. In the multi-access format, the user can begin from any character and move to others as they wish.*

There are two types of identification keys: single-access keys (figure 1.a) and multi-access keys (figure 1.b; 12). In the single-access key model, the reviewer is confronted with a fixed set of identification steps, in a fixed sequential order. Each step in this process is called a node, and presents the reviewer with a set of choices. This set of choices is called a couplet, and the outcome of that choice leads to the next set of decisions. At some point the choice will lead to some final categorization of the reviewed item (12). Decisions are either dichotomous (two outcomes possible) or polytomous (multiple outcomes possible), although the dichotomous option is more often seen. The dichotomous decision set-up is so common that these types of keys are colloquially known as dichotomous keys even though the decision system is not always dichotomous in nature.

Multi-access keys operate in a very different format from single-access keys with the same end result. In this model, the reviewer can approach identification from any step, and follow the next steps in any sequence until a final decision is reached. Multi-access keys are often digital, interactive keys, such as a Lucid key (Lucidcentral.org). Since our work does not involve multi-access keys, we will not elaborate further on this concept.

Identification keys are often tied to the taxonomy (arrangement methods) of the groups being studied. Systematic designation systems, such as the modern Biological Classification System (BCS; 13) or the Soil Taxonomy System (STS; 5) present an all-inclusive tiered system, whereby all lower classification levels fall under a higher order tier. For example, under the modern BCS the major taxonomic categories (taxon) are:

Life>Domain>Kingdom>Phylum>Class>Order>Family>Genus>Species

Mammals and Diptera (flies) are both classified together under the kingdom Animalia, but belong to different phyla (Chordata vs. Arthropoda). Therefore, if we were presented with an identification key that keyed to the level of kingdom only, both *Homo sapiens* (a mammal) and *Phormia regina* (a dipteran) would key out together at that level. Good identification keys include couplets that are diagnostic. A diagnostic couplet is one wherein you can distinguish a single taxon from all others. Not all nodes can be diagnostic. For example, in Figure 1a, node one is non-diagnostic, whereas nodes two and tree are diagnostic, since they result in the smallest-level distinction available on the key (species-level, in this case).

Keys are diagnostic to the smallest level of category they are designed to distinguish. This is true even when the identification key being used features a visual component, since those visual keys work off of an underlying written description, and often include written descriptions of the key distinguishing characters in the couplet with the visual component. Taxon diagnosis requires a broad understanding of the taxa in question. A great deal of exclusive vocabulary can surround the very precise work of taxon diagnosis, especially since taxonomists do not always agree on the importance of features to produce a taxonomy. Authoritative works such as the International Zoological Code of Nomenclature (http://iczn.org/code) aims to stabilize the naming and revision of animal names. At the most basic level of the key, if the feature(s) of the couplet is unable to diagnostically distinguish the taxa, then it is not a fully diagnostic key.

Proper use of identification keys is integral to proper decision making and can have far reaching impacts. For example, the misidentification of a long-horned beetle species in Canada as a native instead of non-native species resulted in a major pest outbreak of this species almost a decade later (14). Likewise, in forensics, the misidentification of a specific drug in a NIK (www.ForensicSource.com) presumptive test could prevent the tester from understanding they have probable cause to further their investigation. For example, it is important to be able to correctly discriminate between morphine, which may be prescribed, and heroin, which is an illicit drug. There is a common understanding (especially among groups such as entomologists) that the use of identification keys requires training and practice in order to become proficient at the use of these tools. Accurate use of identification keys requires training and practice. The first time a person uses a key the results are likely to be poor (14). Anecdotally, undergraduate students in advanced entomology courses that include training on identification keys consistently have error rates as high as 50% even after a full semester of practice (14).

An alternative to using identification keys is relying upon specimen visual gestalt, or "sight ID," for identification. In this type of situation, visual ques are assessed simultaneously without a defined key or methodology. Unknown specimens are compared to known assemblages, and from this comparison, identifications are made. We could find no discussion on the relative success of such endeavors, but this is generally seen as a common decision-making system among many biological systematics groups. There are potential problems with this gestalt approach: 1) without the formal dichotomous keys to guide the user, the decision can be biased by user preferences or knowledge limitations (15), 2) there is a lack of confidence in the results, due to no standard being followed (16).

In recent decades, there has been increasing emphasis on STEM education and student performance in preparation for STEM careers (17). This emphasis on STEM education should not erase the interdisciplinary research and learning that happens in undergraduate education. Interdisciplinary research and interdisciplinary instruction opportunities have been a cornerstone of immersive and intensive experiences for undergraduate students (18). Interdisciplinary instruction and research realize that there are differing skill sets of STEM and Non-STEM majors. A degree from a STEM field, as defined by the U.S. Department of Homeland Security (19), is a degree that contains engineering, biological sciences, mathematics, technology, physical sciences, or a related field. As defined, emphasis on the STEM disciplines may leave out training in dozens of other majors and disciplines that are equally valuable and necessary for the workforce. Forensic Science, classified as a STEM discipline under these criteria, nevertheless contains elements of Non-STEM education that are vital to producing a good forensic scientist. In this way, a more balanced interdisciplinary approach is needed, and these are indeed highlighted in the Forensic Science Education Program Accreditation Commission standards 4.1a (revised 2019; http://fepac-edu.org) when they include topics such as: courtroom testimony, introduction to law, quality assurance, ethics, professional practice, and evidence identification, collection, and processing.

The value of the interdisciplinary instructional and research opportunities has proven invaluable to student success in the educational, research, and lab settings (18). STEM majors are often well-suited for "concrete tasks" with emphasis on tangible objectives with "right answers". Non-STEM majors tend to be more suited on the holistic and non-tangible aspects of project management (20). Successful interdisciplinary teamwork combines the strengths of both sets of cognitive and educational attributes for more comprehensive and productive projects and assignments (21). Furthermore, employers recruit with interdisciplinarity in mind, focused on: technical aptitude, interpersonal skills, and team playing (22).

Interdisciplinary teamwork is productive, efficient, and accurate. Springer, Stanne, and Donovan (23) conducted a meta-analysis of the effects of small group learning on undergraduates and found positive correlations between small team learning in STEM and achievement, persistence, and attitudes. They found that students in small groups performed better than those who were not exposed to group learning (23). In our study, we investigated the performance of small teams (pairs) in a STEM course when pairs are constructed as STEM only, Non-STEM only, and interdisciplinary (STEM + Non-STEM) teams. We explore whether interdisciplinarity also matters in introductory forensic science STEM courses.

In our study, sought to understand how well students successfully followed visual dichotomous keys to the correct identification, with the goal of figuring out how to alter the labs to improve student performance. We hypothesized that 1) students would improve their ability to correctly identify a specimen to species when exposed to the same keys two weeks in a row, 2) when paired with another individual and had to agree on a final identification together, after having worked on that identification first by themselves, 3) students would improve from their initial success rate once paired with another student and required to compare their answers, 4) student self-evaluation of confidence in decisions would help guide them in identifying mistakes. We exposed undergraduate students taking a forensic analysis course to basic insect anatomy and visual species identification using a publicly available online key for forensically important flies (24). Over two lab sessions, students completed identifications for multiple specimens of either adult flies using an online key (24), larval flies using an in-house key (see supplemental S1), or adult carrion beetles (25). In lab one, paired students individually recorded their nodal decisions, rated their confidence in those decisions, and then compared those decisions with their partner, with an option to correct if they found differences. In lab two, individual students recorded their nodal decisions, rated their confidence in those decisions, but did not compare results with a partner.

Methods

Confidence calibration

Calibration refers to the degree to which learners' judgments about their own learning or decisions match the level of learning or decision accuracy they actually manifest (e.g., 26). Thus, students' calibration is a key factor on their self-regulation of learning (e.g., 27). For example, when students are overconfident they are likely to fail using better study strategies or allocate the necessary time to improve performance (e.g., 28). On the other hand, when students are under-confident, they might allocate unnecessary time and effort to learn items that had already been mastered (29). Calibration is particularly important as it is significantly related to academic performance (28).

The level of confidence in one's own answer or decision is also related to the likelihood of correcting that answer or decision. Interestingly, and somewhat counter-intuitively, high confidence errors are more likely to be corrected than errors made with low confidence (29, 30, 31, 32). This effect – hypercorrection of high-confidence errors – is particularly striking because most theories posit that high confidence responses are the ones learners believe most strongly or that have a stronger activation in memory (33) and therefore should be less likely to be updated than low confidence responses. However, to our knowledge, no one has looked at: 1) the relationship between confidence and willingness to change a correct response to a wrong response or 2) effects of confidence in collaborative decision making. Moreover, the effects of confidence in real world classroom activities has also been highly disregarded (for an exception, see 34).

In our studies we accounted for gender and researched the relationship between this variable and both calibration [or confidence level] and likelihood of response change after collaboration. Lundeberg, Fox, & Puncochar (35) reported that undergraduate students are in general overconfident in their exam responses but when they are incorrect, male students are more overconfident than female students. Regarding academic performance, students with higher academic performance tend to be lees overconfident than student with poorer academic performance (27).

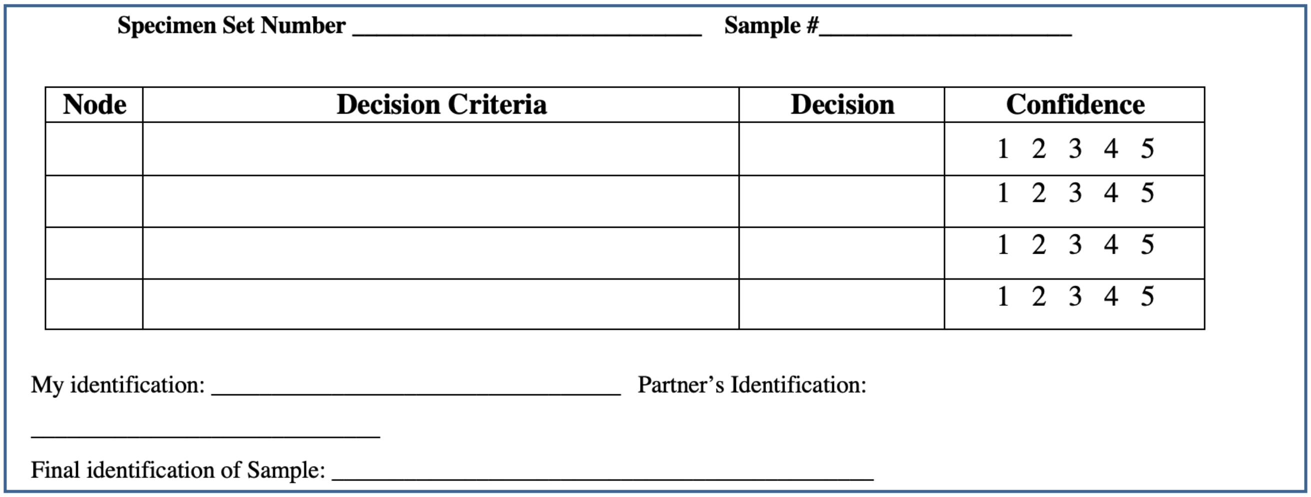
Experiment 1

Subjects

Data were collected from undergraduate students enrolled in ENTM 22820: Forensic Analysis during Spring 2016 (N = 101) and Spring 2017 (N=114). These data came from student participation in regular laboratory activities. This course is open to all disciplines as a part of the core curriculum of the university

Procedure & Materials

During a 110-minute laboratory class, an individual student was asked to identify eight forensically important specimens. Four of the specimens consisted of forensically important blow flies (Diptera: Calliphoridae) and four specimens consisted of forensically important beetles (Coleoptera). Each student was to find a partner and work individually within their pair and compare their answers at the end. Prior to the laboratory class, each student had to label a lateral image of a blow fly, including four directional terms (anterior, posterior, dorsal, and ventral) along with eleven characters found on a blow fly (leg, wing, anterior spiracle, calypters, basicosta, palps, antennae, aristae, meron, halteres and gena). These characters were found with the use of a blow fly dichotomous pictorial key (24). In the 2017 collection period, care was taken in class to ensure students completely filled out their identification worksheets, especially to provide confidence ratings, which were not always recorded in experiment 1.



Specimen Set Number ______________________________ **Sample #**_____________________

Node	Decision Criteria	Decision	Confidence
			1 2 3 4 5
			1 2 3 4 5
			1 2 3 4 5
			1 2 3 4 5

My identification: ___________________________________ Partner's Identification: ______________________________

Final identification of Sample: ____

FIGURE 2: Depiction of Table used during identification by students.

Blow fly specimens

Each student identified the four blow fly specimens alone with the use of a Leica EZ4 HD stereomicroscope, using the Cutter & Dahlem (24) blow fly key. Each specimen contained an identifier associated with a key, which gave no indication of species. Students were instructed to begin under the heading entitled "Identification of Calliphoridae Species". The Cutter & Dahlem (24) key is a dichotomous pictorial key, allowing each student to focus on a particular trait and decide between one of two options, i.e. is a character absent or present? As the student selected their choice it brought them to another node, where they again repeated their choice selection on a new character. This continued until they reached an end point, which in this key would be a species identification. Each student decision was recorded and students assigned a confidence level (1-5) as shown in figure 2. A value of 1 indicated a low confidence in their decision, while a 5 indicated a high confidence in their decision. Along with their decision criteria each student recorded the specimen number and their own identification. After each person within a pair completed the identification by themselves, they compared their answers to their partner (also recording this information on their worksheet). If both partner's identifications matched they recorded the species name under "final identification of sample", if the specimens did not match they reanalyzed their decision table to see where they differed. They then reanalyzed the specimen to see if they reached an identification they both agreed upon and recorded their final answer on the appropriate line.

Beetle specimens

Students were provided with a "Forensic Insect Identification Cards" pictorial flipbook (25) and four unknown beetle specimens. As with the blow flies, each beetle contained an identifier associated with a key, which gave no indication of species. Each student recorded the specimen set and sample number of each beetle. They then looked through the flip book until they found the beetle that they thought was the specimen they had. They had to write a justification as to why they chose that beetle species and record the length of each beetle. After each person in a pair completed the identification by themselves, they compared their answers to their partner (also recording this information on their worksheet).

Data alignment and analysis

Nodal Decisions

The Cutter & Dahlem (24) and the Internal Larval Key (S1) dichotomous keys are organized so that, when starting at the beginning of the key, decisions either direct to the next node or provide species identification. Student nodal decisions were recorded in numbered format according to the next position the key sent them to, until they reach the identification that was recorded as "ID". Some students began their identification at a key to families of Diptera before reaching the key to species of Calliphoridae (24), which added multiple steps to the identification process and therefore made it more difficult for them. These students' responses were recorded but not used in the analysis. The identification of beetles did not use a dichotomous key, but instead a set of Forensic Insect Identification Cards (25), therefore it was not possible to record their nodal decisions, but their initial and final identification was recorded using a format similar to Figure 1.

Confidence

Along with their nodal decisions students were asked to rate their confidence at each decision they made on a scale from 1 (not confident at all) to 5 (very confident). This was recorded using the same numbered scale from 1-5. If a student did not rate their confidence at any step, a period was recorded into the spread sheet to show that information was not available, and they were dropped from the analysis.

Data analysis

Performance data were analyzed, in terms of accuracy, for each individual student and also for the pair. Performance and confidence differences between majors (STEM vs. Non-STEM) were also analyzed. Only data from students who worked in pairs and completed all the confidence ratings were included in the analyses. Differences in performance between the two keys were analyzed using paired samples *t*-tests. Differences in performance and confidence depending on majors were analyzed using independent samples *t*-tests. Pairs constitution and performance was subject to descriptive analysis. Performance in Lab 1 and Lab 2 was compared using paired samples *t*-tests. Finally, the average accuracy of students' decision in each node in the key for fly identification (in the adult stage for both laboratories, and in the larvae stage for Laboratory 2) was calculated. This nodal accuracy was subjected to a descriptive analysis, by species. For all the tests performed, the assumptions for constant variance and normality were checked with preliminary diagnostic analyses. Whenever the equal variance assumption was violated, a Welch's *t*-test was performed instead of the standard *t*-test. Likewise, whenever the normality assumption was violated, a nonparametric test was performed to compare the group differences. Wilcoxon Sign-Rank test was used in place of paired *t*-test, and Mann Whitney Wilcoxon test was chosen as an alternative to independent samples *t*-test. All performance data are presented in terms of proportions (where 1 would equal 100% accuracy) and all confidence data are presented in terms of averages of values from the 1-5 scale used.

Table 1—*Mean performance and confidence ratings, per semester, laboratory, and type of identification (standard deviations in parenthesis).*

Cohort	2016 Lab 1	2017 Lab 1	2016 Lab 2	2017 Lab 2
Flies Initial Accuracy	0.39 (0.27)	0.36 (0.26)	0.45 (0.32)	0.45 (0.26)
Flies Final Accuracy	0.41 (0.27)	0.36 (0.27)		
Beetles Initial Accuracy	0.96 (0.09)	0.84 (0.26)		
Beetles Final Accuracy	0.93 (0.25)	0.83 (0.29)		
Flies Confidence	4.29 (0.57)	3.81 (0.62)	4.15 (0.56)	3.94 (0.55)
Larval Fly Stages Accuracy			0.51 (0.23)	0.43 (0.26)
Larval Fly Stages Confidence			4.29 (0.49)	4.11 (0.55)

Results

Laboratory 1

A total of 32 students' data were analyzed in 2016 (STEM =12, Non-STEM = 20), and this was increased to one hundred students in 2017 (STEM =36, Non-STEM = 64). Regarding performance in species identification (see Table 1 for the full data on performance and confidence for Spring 2016 and 2017, Lab 1), students' initial accuracy did not differ from their final accuracy, i.e., after discussing with the partner. This pattern occurred for both the flies identification (2016: *M* = .39 vs. *M* = .41; 2017: *M* = .36 vs. *M* = .37, for initial and final accuracy, respectively), 2016: *t*(31) = 0.37, *p* = .712, *d* = .06, 2017: *t*(99) = 0.52, *p* = .604, *d* = 0.03; and for beetles identification (2016: *M* = .96 vs. *M* = .94; 2017: *M* = .84 vs. *M* = .83, for initial and final accuracy, respectively), 2016: Wilcoxon Sign-Rank *Z* = -0.41, *p* = .679, 2017: *t*(99) = -0.89, *p* = .374, *d* = -0.03. The comparison between the accuracy for each one of the keys showed significantly higher accuracy for the beetles identification than for the flies identification, both for the initial identification, 2016: Wilcoxon Sign-Rank *Z* = 4.77, *p* < .001; 2017: *t*(99) = 11.61, *p* <.001, *d* = 1.74, and for the final identification, *2016:* Wilcoxon Sign-Rank *Z* = 4.83, *p* < .001; 2017: *t*(99) = 11.28, *p* <.001, *d* = 1.62.

Differences between majors were also analyzed (see Table 2). Overall, no significant effects emerged. Both STEM and non-STEM majors showed similar initial accuracy when identifying flies (2016: *M* = .46 vs. *M* = .35; 2017: *M* = .37 vs. *M* = .35, for STEM and non-STEM majors, respectively), 2016: *t*(30) = 1.11, *p* = .277, *d* =.41; 2017: *t*(98) = 0.30, *p* = .766, *d* = 0.06, and when identifying beetles (2016: *M* = .96 vs. *M* = .96; 2017: *M* = .89 vs. *M* = .80, for STEM and non-STEM majors, respectively), 2016: Mann Whitney Wilcoxon *Z* = -0.12, *p* = .902; 2017: Welch's *t*(96.56) = 1.61, *p* = .110, *d* = 0.31. The same pattern was obtained for the final accuracy for flies identification (2016: *M* = .0.42 vs. *M* = .40; 2017: *M* = .35 vs. *M* = .37, for STEM and non-STEM majors, respectively), 2016: *t*(30) = 0.17, *p* = .868, *d* = 0.06; 2017: *t*(98) = -0.30, *p* = .767, *d* = 0.06. For the final accuracy for beetles identification, a numerical advantage for non-STEM students emerged for 2016, while a numerical advantage for STEM students emerged for 2017 (2016: *M* = .83 vs. *M* = 1.00; 2017: *M* = .89 vs. *M* = .79, for STEM and non-STEM majors, respectively). However, neither reached significance, 2016: Mann Whitney Wilcoxon *Z* = -1.86, *p* = .063; 2017: Welch's *t*(97.27) = 1.87, *p* = .064,

Table 2—*Accuracy and confidence for STEM and Not-STEM majors (Means are presented and standard deviations are in parenthesis).*

	Group	2016 - Lab 1	2017 - Lab 1	2016 - Lab 2	2017 - Lab 2
Flies Initial Accuracy	Non-STEM	0.35 (0.27)	0.35 (0.28)	0.41 (0.30)	0.43 (0.27)
	STEM	0.46 (0.26)	0.37 (0.23)	0.59 (0.39)	0.48 (0.25)
Flies Final Accuracy	Non-STEM	0.40 (0.25)	0.37 (0.28)		
	STEM	0.42 (0.31)	0.35 (0.25)		
Beetles Initial Accuracy	Non-STEM	0.96 (0.09)	0.80 (0.32)		
	STEM	0.96 (0.10)	0.89 (0.20)		
Beetles Final Accuracy	Non-STEM	1.00 (0.00)	0.79 (0.33)		
	STEM	0.83 (0.39)	0.89 (0.20)		
Flies Average Confidence	Non-STEM	4.23 (0.57)	3.75 (0.63)	4.14 (0.53)	3.95 (0.56)
	STEM	4.39 (0.58)	3.93 (0.57)	4.22 (0.71)	3.93 (0.55)
Larval Fly Stages Accuracy	Non-STEM			0.50 (0.23)	0.47 (0.27)
	STEM			0.57 (0.23)	0.37 (0.21)
Larval Fly Stages Confidence	Non-STEM			4.26 (0.50)	4.09 (0.60)
	STEM			4.43 (0.46)	4.12 (0.48)

d= 0.36. Average confidence[1] ratings in flies' identifications also did not differ between majors (2016: *M* = 4.39 vs. *M* = 4.23; 2017: *M* = 3.93 vs. *M* = 3.75, for STEM and non-STEM majors, respectively), 2016: *t*(30) = 0.759, *p* = .454, *d* = 0.28; 2017: *t*(98) = 1.40, *p* = .164, *d* = 0.30.

Pair performance (i.e., final accuracy), depending on their constitution was also analyzed. In 2016, there was a total of 16 pairs, eight of which were constituted by one STEM major and one non-STEM major (mixed pairs), six of which were constituted by two non-STEM majors (non-STEM pairs), and by two STEM majors (STEM pairs). In 2017 there was a total of 50 pairs, 16 of which were constituted by one STEM major and one non-STEM major (mixed pairs), 24 of which were constituted by two non-STEM majors (non-STEM pairs), and ten by two STEM majors (STEM pairs). Given the low number of pairs in 2016, no statistical analyses were performed, although a description of the data is depicted in Table 3. For 2017, independent samples *t*-tests were performed on final accuracy for fly identifications and on final accuracy for beetle identifications. Regarding final accuracy on flies identification, mixed pairs performed better than non-STEM pairs (*M* = .45 vs. *M* = .29, respectively), although the differences did not reach significance, *t*(38) = 1.82, *p* = .076, *d* = 0.59. Mixed pairs and STEM pairs did not differ greatly (*M* = .45 vs. *M* = .40, respectively), despite a numerical advantage for mixed pairs, *t*(24) = 0.54, *p* = .594, *d* = 0.22. This pattern of results was the same for beetle identifications. Mixed pairs performed better than non-STEM pairs (*M* = .92 vs. *M* = .71, respectively), Welch's *t*(36.82) = 2.17, *p* = .037, *d* = 0.66. Mixed pairs and STEM pairs did not statistically differ (*M* = .92 vs. *M* = .85, respectively), despite a numerical advantage for mixed pairs, *t*(24) = 0.78, *p* = .440, *d* = 0.31.

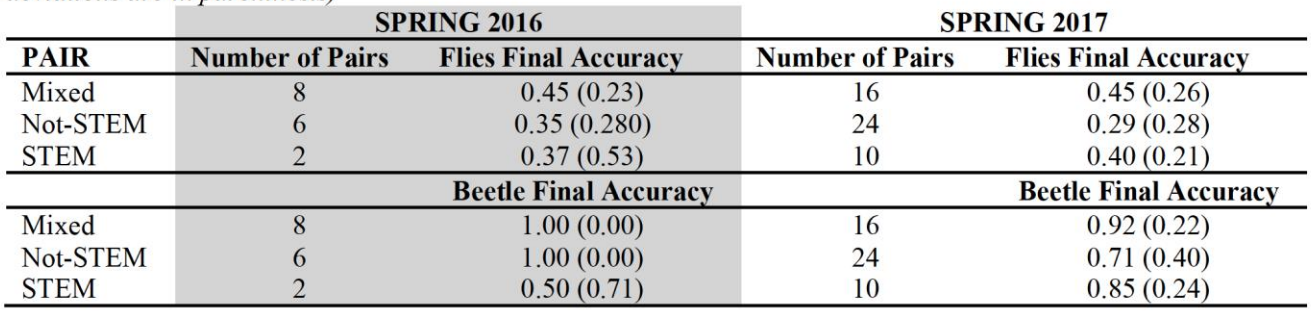
Laboratory 2

A total of 57 and 84 students' data were analyzed for 2016 and 2017, respectively. Regarding performance (see Table1), students were equally accurate when identifying adult flies (2016: *M* = .45; 2017: *M* = .45) and larval fly stages (2016: *M* = .51; 2017: *M* = .43), 2016: *t*(56) = 1.48, *p* = .145, *d* = 0.24; 2017: *t*(83) = -0.53, *p* = .599, *d* = -0.06. However, students were significantly less confident in their accuracy when identifying adult flies (2016: *M* = 4.15; 2017: *M* = 3.94) than when identifying larval fly stages (2016: *M* = 4.29; 2017: *M* = 4.11), 2016: *t*(56) = 2.90, *p* = .005, *d* = 0.26; 2017: *t*(83) = 3.86, *p* <.001, *d* = 0.29. Similarly, to what happened in Lab 1, no significant major differences were obtained, although it should be mentioned that of the 50 students who reported their major, only 11 were STEM majors in 2016.

Comparison between Laboratory 1 and Laboratory 2

Accuracy in adult flies' identification was compared between laboratories (Lab 1 vs. Lab 2; see Table 4). A total of 108 students' data (2016: *n* = 32; 2017: *n* = 76) were analyzed, using a paired samples *t*-test. Initial students' classifications were not significantly different between Lab 1 and Lab 2 in 2016 (M_{lab1} = 0.39, $M_{\text{lab 2}}$ = 0.41, *t*(31) = 0.21, *p* = .840, *d* = 0.05), whereas initial students' classifications were more accurate in Lab 2 than in Lab 1 in 2017($M_{\text{lab 1}}$ = .37 vs. $M_{\text{lab 2}}$ = .45, *t*(75) = 2.31, *p* = .024, *d* = 0.34. Regarding the comparison between final students' classifications in Lab 1 and individual classification in Lab 2, no significant differences were found in 2016 (M_{lab1} = 0.41, $M_{\text{lab 2}}$ = 0.41, *t*(31) = 0.00, *p* = 1.00, *d* = 0.00) and 2017 ($M_{\text{lab 1}}$ = .37 vs. $M_{\text{lab 2}}$ = .45, *t*(75) = 1.96, *p* = .054, *d* = 0.30), although a numerical advantage for Lab 2 emerged in 2017. There were no significant differences between confidence ratings in Lab 1 and Lab 2 in 2016 (*M* = 4.29 vs. *M* = 4.15, respectively), *t*(31) = 1.04, *p* = .307, *d* = 0.25. While in 2017, there were significant differences

Table 3—*Pair performance on Flies and Beetles Identification, by type of pair (Means are presented and standard deviations are in parenthesis)*

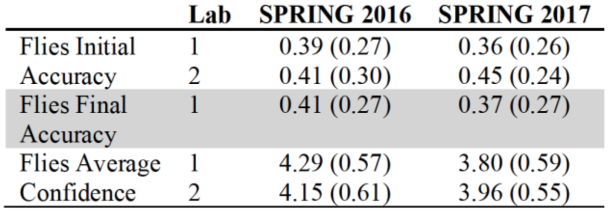


	SPRING 2016		SPRING 2017	
PAIR	**Number of Pairs**	**Flies Final Accuracy**	**Number of Pairs**	**Flies Final Accuracy**
Mixed	8	0.45 (0.23)	16	0.45 (0.26)
Not-STEM	6	0.35 (0.280)	24	0.29 (0.28)
STEM	2	0.37 (0.53)	10	0.40 (0.21)
		Beetle Final Accuracy		**Beetle Final Accuracy**
Mixed	8	1.00 (0.00)	16	0.92 (0.22)
Not-STEM	6	1.00 (0.00)	24	0.71 (0.40)
STEM	2	0.50 (0.71)	10	0.85 (0.24)

[1] Average confidence was calculated by averaging each nodal confidence rating for each one of the species identified.

between confidence ratings in Lab 1 and Lab 2, with students being less confident in Lab 1 (*M* = 3.80) than in Lab 2 (*M* = 3.96), *t*(75) = 3.07, *p* = .003, *d* = 0.28.

Table 4—*Performance and confidence ratings in Lab 1 and Lab 2, for students who performed all the tasks in both labs (Means are presented and standard deviations are in parenthesis).*



	Lab	SPRING 2016	SPRING 2017
Flies Initial Accuracy	1	0.39 (0.27)	0.36 (0.26)
	2	0.41 (0.30)	0.45 (0.24)
Flies Final Accuracy	1	0.41 (0.27)	0.37 (0.27)
Flies Average Confidence	1	4.29 (0.57)	3.80 (0.59)
	2	4.15 (0.61)	3.96 (0.55)

Discussion and Conclusion

It is generally thought that students begin poorly but improve upon repeated use of dichotomous keys. Our findings support this claim. When compared individually, there appear to be no difference in the initial to final accuracy within a single identification event, regardless of whether a key is used (in the case of flies) or a gestalt visual identification system (beetles; Table 1). Individual student accuracy does increase with repeated exposure to the key, as shown here by increases in initial accuracy in lab 2 over initial and final accuracy in lab 1 (Table 1). STEM and Non-STEM students perform equally well with initial decision accuracy for both flies and beetles, but STEM students increase accuracy over Non-STEM in final accuracy (Table 2). However, when students are classified into STEM/STEM, STEM/Non-STEM (mixed) and Non-STEM/Non-STEM pairings, we see significant gains in the mixed pairings over either of the pure pairings (Table 3). This supports the earlier research of Springer et al. (23) where STEM teams outperformed non-STEM teams, and builds off it, indicating that mixed teams perform even better. Interestingly, we find no difference in gender, amongst these gains, and no relationship between individual student confidence and success (results not reported).

The results are potentially profound—mixing STEM Non-STEM students seems to produce the best gains when practicing this type of skill. The gains are statistically significant, and require little effort on the part of the educator to put into practice. Organizing student groups to maximize STEM / Non-STEM relationships should be relatively easy in a modern school setting, where majors are knowable and databased. We hypothesize that these increased results are based upon student interactions, and the unique perspectives they bring to the classroom. However, this requires further research to verify.

There is a potential confounding problem between the 2016 and 2017 data. In 2016 we discovered that students did not always record their nodal decisions, something that confounds our ability to investigate how students influence each other's decision-making capabilities. In 2017 TAs were instructed to specifically work with students to ensure they were properly filling out their nodel decisions prior to comparison—as the lab was intended to be completed. Because of this, only 32 student's data from 2016 was usable by our analysis, compared to 100 students data for 2017. This means that for laboratory 1 only 31.68% of 2016 samples were evaluated, with the remaining being removed due to incomplete data or not working in pairs, but in 2017 this increased to an 87.72% acceptance rate. It is possible that this injected some sort of bias or other issue into the evaluation. Because of this, we recommend that further studies should be undertaken to confirm these results.

From a forensics viewpoint, this seems to indicate that introductory courses might not be best served by being "majors only" but rather can be organized for the largest learning gains by being diverse in nature. Further work should be carried out to verify that this type of learning trend continues when student populations are mixed, both for dichotomous keys and the other areas of forensic science referenced in the introduction of this article (e.g.—NIK kit testing, anthropology, etc.).

Acknowledgements

The authors wish to thank Gregory Dahlem and Regina Cutter for producing their online visual key. Such work allows for basic instruction in these essential tools of the naturalist. Further, TS thanks all his students who took the ENTM 22820 course, and also the student and graduate student teaching assistants who ensured uniformity amongst lab sections. Without their work, this analysis would not be possible.

References

1. Castro-Valderrama, U., Carvalho, G. S., Peck, D. C., Valdez-Carrasco, J. M., & Napoles, J. R. (2019). Two New Species of the Spittlebug Genus Ocoaxo Fennah (Hemiptera: Cercopidae) from Mexico, and Keys for the Groups, Group Three, and First Subgroup. *Neotropical Entomology, 48*(2), 260-268. doi:10.1007/s13744-018-0629-0
2. Contico, F. P., & Fleischmann, A. (2016). The first record of the boreal bog species *Drosera rotundifolia* (Droseraceae) from the Philippines, and a key to the Philippine sundews. *Blumea, 61*(1), 24-28. doi:10.3767/000651916x691330
3. Lacourse, T., Beer, K. W., & Hoffman, E. H. (2018). Identification of conifer stomata in pollen samples from western North America (vol 232, pg 140, 2016). *Review of Palaeobotany and Palynology, 258*, 265-265. doi:10.1016/j.revpalbo.2018.07.003

4. Gillung, J. P., & Winterton, S. L. (2011). New genera of philopotine spider flies (Diptera, Acroceridae) with a key to living and fossil genera. *Zookeys* (127), 15-27. doi:10.3897/zookeys.127.1824
5. Staff, S. S. (1999). *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys, Second Edition*. (436). National Resource Conservation Service.
6. Zanella, A., Schad, P., Galbraith, J., & Ponge, J. F. (2018). Humusica 2, Article 14: Anthropogenic soils and humus systems, comparing classification systems. *Applied Soil Ecology, 122*, 200-203. doi:10.1016/j.apsoil.2017.07.006
7. Alan Plante, Donald Peck and David Von Bargen. (2003) *Mineral Identification Key*. Mineralogical Society of America.
8. Galton, F. (1892). *Finger Prints*. London: MacMillan and Co.
9. Trail, P. W. (2017). Identifying Bald Versus Golden Eagle Bones: A Primer for Wildlife Biologists and Law Enforcement Officers. *Journal of Fish and Wildlife Management*, *8*(2), 596-610. doi:10.3996/042017-jfwm-035
10. Whitworth, T. (2006). Keys to the genera and species of blow flies (Diptera: Calliphoridae) of America north of Mexico. *Proc. Entomol. Soc. Wash., 108*(3), 689-702.
11. Symonsbergen, D. J., Kangas, M. J., Perez, M., & Homes, A. E. (2018). Evaluation of the NIK® test: Primary general screening test for the presumptive identification of drugs. *International Journal of Criminal and Forensic Science, 2*(5), 81-137.
12. Quicke, D. L. J. (1993). *Principles and Techniques of Contemporary Taxonomy*: Blackie Academic & Professional.
13. Mayr, E. (1942). *Systematics and the origin of species from the viewpoint of a zoologist*. New York: Columbia University Press.
14. Marshall, S. A. (2000). Comments on error rates in insect identifications. *Newsletter of the Biological Survey of Canada (Terrestrial Arthropods)*. Retrieved from http://www.biology.ualberta.ca/bsc/news_19_2/error_rates.htm
15. Winston, J.E. (1999). Describing species: practical taxonomic procedure for biologists. Columbia University Press, New York, USA.
16. Strengthening Forensic Science in the United States: A Path Forward. (2009). National Academies Press, Washington, D.C., USA.
17. Olson, S., & Riordan, D. G. (2012). Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. Report to the President. *Executive Office of the President*.
18. Borrego, M. & Newswander,L.K. (2010). Definitions of Interdisciplinary Research: Toward Graduate-Level Interdisciplinary Learning Outcomes. The Review of Higher Education 34(1), 61-84. Johns Hopkins University Press. Retrieved June 4, 2019, from Project MUSE database.
19. "STEM designated degree program list." Department of Homeland Security. Accessed April 7th, 2020: https://studyinthestates.dhs.gov/2016/05/how-does-dhs-determine-which-degrees-qualify-for-the-stem-opt-extension
20. Capraro, R. M., Capraro, M. M., & Morgan, J. (Eds.). (2013). *Stem project-based learning: An integrated science, technology, engineering, and mathematics (stem) approach*. Retrieved from https://ebookcentral.proquest.com
21. Ferrari, N., Jenkins, C., Garofano, J., Day, D., Schwendemann, T., & Broadbridge, C. (2015). Research Experiences for Students: Interdisciplinary skill development to prepare the future workforce for success. MRS Proceedings, 1762, Mrsf14-1762-aaa08-03. doi:10.1557/opl.2015.154
22. Vilorio, Dennis. (2014). STEM 101: Intro to tomorrow's jobs. (science, technology, engineering and mathematics). Occupational Outlook Quarterly, 58(1), 2-12.
23. Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of educational research*, *69*(1), 21-51.
24. Cutter, R. M., and G. Dahlem. 2004. *Identification key to the common forensically important adult flies (Diptera) of northern Kentucky*. Department of Biological Sciences, Northern Kentucky University, Highland Heights, KY.
25. Castner, J. L., & Byrd, J. H. (2000). *Forensic Insect Identification Cards*. China: Feline Press.
26. Lichtenstein, S., B. Fischhoff, L. D. Phillips. 1982. Calibration of probabilities: The state of the art to 1980. D. Kahneman, P. Slovic, A. Tversky, eds. Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, UK, 306–334.
27. Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated engagement in learning. in metacognition in educational theory and practice. Metacognition in Educational Theory and Practice, 277-304.
28. Hadwin, A. F., & Webster, E. A. (2013). Calibration in goal setting: Examining the nature of judgments of confidence. *Learning and Instruction*, *24*, 37-47.
29. Metcalfe, J., & Finn, B. (2011). People's hypercorrection of high-confidence errors: Did they know it all along? *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, *37*(2), 437.

30. Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1491.
31. Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning*, *1*(1), 69-84.
32. Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604-616.
33. Koriat, A. (1997). Monitoring One's Own Knowledge during Study: A Cue-Utilization Approach to Judgments of Learning. *Journal of Experimental Psychology: General*, 126, 349-370. http://dx.doi.org/10.1037/0096-3445.126.4.349
34. Alexander, P. (Ed.) (2013) Calibrating Calibration: Creating Conceptual Clarity to Guide Measurement and Calculation [special issue]. *Learning and Instruction*, 24, 1-66.
35. Lundeberg, M. A, Fox, P. W., & Puncochar, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology*, 86,114-121.