A card-sorting tool to measure expert versus novice thinking in scientific research

Megan F. Cole,[†]* Clarke O. Britton,[‡] Denver Roberts,[†] Peter Rubin,[†] Hannah D. Shin,[‡] Yassin R. Watson,[‡] and Colin Harrison[‡]*

[†]Department of Biology, Emory University, Atlanta, GA 30322; [‡]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332

ABSTRACT

Undergraduate research and laboratory experiences provide a wide range of benefits to student learning in science and are integral to imbed authentic research experiences in biology labs. While the benefit of courses with research experience is widely accepted, it can be challenging to measure conceptual research skills in a quick and easily scalable manner. We developed a card-sorting task to differentiate between novice and expert conceptualization of research principles. There were significant differences in the way faculty/postdocs, graduate students, and undergraduate students organized their information, with faculty/postdocs more likely to use deep feature sorting patterns related to research approach. When provided scaffolding of group names reflecting expert-like organization, participant groups were better able to sort by that organization, but undergraduate students did not reach expert levels. Undergraduates with Advanced Placement experience were more likely to display expert-like thinking than undergraduates without Advanced Placement Biology experience and non-PEER (persons excluded because of their Ethnicity or Race) students displayed more expert-like thinking than PEER students. We found evidence of undergraduates in various stages of development toward expert-like thinking in written responses. This card-sorting task can provide a framework for analyzing student's conceptualizations of research and identify areas to provide added scaffolding to help shift from novice-like to expert-like thinking.

INTRODUCTION

Increasing use of Course-Based Undergraduate Research Experiences (CUREs) in college curriculums is allowing more students to gain firsthand experience with research (Bangera and Brownell, 2014; Corwin Auchincloss *et al.*, 2014; Shortlidge *et al.*, 2016). Past work has shown that providing students with this early experience can yield many benefits such as fueling a desire to obtain advanced degrees or enter careers in Science, Technology, Engineering, and Mathematics (STEM; Russell *et al.*, 2007; Harrison *et al.*, 2011; Graham *et al.*, 2013; Rodenbusch *et al.*, 2016), boosting confidence in students' scientific abilities and understanding, and increasing sense of belonging in science (Seymour *et al.*, 2004; Kuh, 2008; Robnett *et al.*, 2015; Hernandez *et al.*, 2020). Ultimately, the goal of CUREs and research-like experiences is to improve students' research skills and expert-like scientific thinking.

Expertise can be described by a number of characteristics that separate expert and novice thinking. One obvious difference is in depth of knowledge where novices tend to rely more on memorization whereas experts have a vast network of interconnected knowledge, allowing them to apply their understanding to complex scientific problems (Chi *et al.*, 1981). Experts are also better able to integrate new information into their existing knowledge, whereas novices can struggle with cognitive overload when acquiring new information (Sweller, 1988; Dreyfus and Dreyfus, 2005). Experts also approach problem solving differently from novices where novices tend to use trial and

Erika Offerdahl, *Monitoring Editor*

Submitted Nov 16, 2022; Revised Jun 20, 2023; Accepted Jul 27, 2023

CBE Life Sci Educ December 1, 2023 22:ar38

DOI:10.1187/cbe.22-11-0230

*Address correspondence to: Megan F. Cole and Colin Harrison (mfcole@emory.edu and colin.harrison@biosci.gatech.edu).

error strategies and struggle to make connections between concepts while experts rely on mental modeling and hypothesis testing to develop new theories and solve complex problems (Chi *et al.*, 1982). Two additional critical differences between experts and novices across domains are experts' ability to recognize meaningful patterns and to "see" problems at a deeper or more principled level than novices, who tend to view problems and patterns on more superficial similarities (Chi *et al.*, 1988; Ericsson & Smith, 1991).

Experts' ability to perceive deep principles connecting concepts or problems leads to a difference in how experts and novices mentally organize and access information (Chi *et al.*, 1981; Bedard and Chi, 1991; Smith, 1992; Krieter *et al.*, 2016; Hoskinson *et al.*, 2017; Galloway *et al.*, 2018). In fact, template theory suggests that by scaffolding students with an expert-like organization structure to use when acquiring new knowledge, educators can better support development of expertise in students (Hoskinson *et al.*, 2017). Once in place, expert-like mental models allow experts to more quickly recognize patterns, retrieve information, and respond appropriately while problem solving (Ericsson and Smith, 1991; Dreyfus, 2004).

The ability to measure expertise is of interest to both better understand experts' mental models and to better support and assess interventions aimed at enhancing or building expertise. Common methods for assessing expertise include observations, interviews, accomplishments, reflections, and knowledge-based assessments. For example, assessing expertise in chess can be accomplished by rankings based on game wins and expertise in medicine may be measured by certification exams. Measuring expertise in students is often tied to specific courses via assignments and exams or, for longer-term progressions, tied to portfolios and concept inventories. Several assessment tools have been developed to measure research-related expertise in students such as the Biological Experimental Design Concept Inventory, Experimental Design Ability Test, Test of Scientific Literacy Skills, Classroom Test of Scientific Reasoning, and California Critical Thinking Skills Test (Facione, 1991; Lederman *et al.*, 2002; Sirum and Humburg, 2011; Gormally *et al.*, 2012; Deane *et al.*, 2014). As these assessments largely measure specific conceptual skills, there is a need for a tool that could measure the underlying shift in how students organize conceptual information as they progress towards expertise. Such a tool may provide insight into how experts versus novices in research organize information and may allow educators to better design and target interventions to assist novices. Additionally, an assessment tool aimed at organization of knowledge should be widely applicable to CUREs with different model organisms, fields of biology, and skills focus.

The difference between how experts and novices perceive and mentally organize information has previously been leveraged in classrooms to quantify expert versus novice-like thinking through use of card-sorting tasks. These tasks involve sorting a set of cards with different concepts, problems, or scenarios into categories based on their relationships. Card-sorting tasks were first used in science to measure differences in how experts and novices sort physics problems (Chi *et al.*, 1981). This study found that experts organize cards based on underlying physics principles used to solve the problems (e.g., conservation of energy), whereas novices focus more on surface similarities between cards (e.g., the type of object involved such as an inclined plane; Chi *et al.*, 1981). This phenomenon of experts sorting based on deep principals and novices sorting based on surface features has been repeated in several studies and fields of science (Lin and Singh, 2010; Mason and Singh, 2011; Smith *et al.*, 2013; Irby *et al.*, 2016; Krieter *et al.*, 2016).

Card-sorting tasks have also been applied to the field of biology. Smith *et al.*, (2013) asked nonmajors Introductory Biology students (novices) and biology faculty (experts) to sort biology problems based on underlying principles. They found that novices and experts used distinct conceptual frameworks to organize the cards; novices sorted based on a superficial framework of phylogenetic group whereas experts sorted based on deep conceptual principles such as evolution by natural selection. This same task was further able to distinguish lower and upper level undergraduate biology students, suggesting that students move towards more expert-like thinking as they progress in a biology major (Bissonnette *et al.*, 2017). A genetics card-sort task was similarly used with genetics undergraduate students, genetics faculty, and genetics counselors but unexpectedly found that the two expert groups (faculty and counselors) used different frameworks to sort the cards with faculty organizing based on conceptual principles and counselors organizing based on problem-solving techniques (Smith, 1992).

Here, we examine whether a card-sorting task could be used to measure differences between novice-, developing-expert-, and expert-like thinking in the field of biology research. Research-based lab courses provide training and experience in research to undergraduate students but assessing student progression towards research expertise can be challenging. Many published CURE assessments focus on either students' perceptions of their experiences and abilities, measuring specific skills or concepts, or are time intensive to score (Stiggins, 2004; Duckworth and Quinn, 2009; Creamer *et al.*, 2010; Sirum and Humburg, 2011; Deane *et al.*, 2014; Lopatto *et al.*, 2014; Makarevitch *et al.*, 2015; Hanauer *et al.*, 2016; Zelaya *et al.*. 2022). Card-sorting tasks have the potential to provide a generalizable and scalable assessment tool that could be used across diverse laboratory course structures, research projects, and student level to measure expert-like thinking in research. We additionally explore whether we observe differences between undergraduates based on Advanced Placement Biology experience, gender, or race.

METHODS

Card-Sorting Task Development

Cards were designed by identifying one deep-level (research approach) and two surface-level (organism studied, and person doing the research) features, then implementing scenarios around each feature. Deep-level research approach categories were anecdotal/story (i.e., flawed research approach or nonresearch), correlational/observational, experimental/manipulative, and secondary/meta-analysis. Surface-level organism categories were humans, fruit flies, plants, and microbes. Surface-level researcher categories were scientist, student, medical professional, and layperson. Each card was assigned to one category for the deep feature, organism-surface feature, and researcher-surface feature (Supplemental Material [card sorting cards and hypothesized sort]). For example, a card assigned to correlational/observational for the deep feature, human for the surface-level organism, and medical professional

for the surface-level researcher reads: "A doctor is interested in the effects of diet on health. They surveyed 2000 individuals in the United States and found that people who ate a fish-based diet had a 10% lower risk for heart disease."

Cards were edited for clarity and length followed by trial runs carried out with populations of experts and novices. Trial run participants performed a sequential sorting of cards, first with only the guidance to sort based on "common underlying scientific principles" (unframed) and then with given category names based on the deep-feature sort (framed). In the expert-level trial run, sorting data and feedback were obtained in a small focus group session of faculty. The expert-level trial run found initial trends consistent with our hypothesized deep-feature sort. Based on feedback, cards and task instructions were edited for clarity of wording. Subsequently, a novice card-sort trial was carried out using 42 first-semester Introductory Biology students. Upon completion of this task, card-sorting instructions were further edited for clarity. Both of these trials were indicative of hypothesized results and this process of trial runs provided an initial face-validity test for our task. The finalized card-sorting task was shifted to an online platform (Qualtrics and Flippity) before final data collection. Final cards can be viewed in supplemental material [card sorting cards and hypothesized sort].

Participant Recruitment and Participant Population

Participants were recruited from two highly selective research universities in the southeast United States; Emory University, a private institution with roughly 7000 undergraduate students, and Georgia Institute of Technology, a state institution with roughly 17,000 undergraduates.

Undergraduate students were recruited from second-semester Introductory Biology labs at Emory and from first- and second-semester Introductory Biology labs at Georgia Tech. For Emory students, the card-sorting task was assigned as an asynchronous assignment worth less than 1% of their overall score for the course and an alternative assignment (of similar effort) was provided as an option for students who did not wish to participate in the study. At Georgia Tech, students were provided a 1% extra-credit opportunity for completing the task outside of class time, or an alternative assignment, for extra credit. The total undergraduate invited pool was 854 students.

Faculty, postdoc, and graduate students were recruited via email from the biology graduate student, postdoc, and faculty populations at both institutions. No incentive for participation was offered to this population and participation rates were low. In total, 658 undergraduate students, 10 graduate students, one postdoc, and 10 faculty responses were collected. Responses were excluded from further analyses when the sorting task was not completed or the subject did not complete the written responses, resulting in a total of 569 undergraduate students, 10 graduate students, and 11 faculty/postdoc responses. Studies were done with approval by IRB (H19330 and 00002179).

The majority of undergraduate students (85%) were either first- or second-year students in college, 65% identified as female, and 27% were from a PEER group (Black/African American, Latinx/Chicano, Native American, Hawaiian/Pacific Islander, or mixed race with at least one of these groups). A majority (77%) of novice participants had no prior research experience. (Table 1)

Survey Instrument

Participants completed the card-sorting task online through a Qualtrics survey and use of an online card-visualization and organization tool (Flippity). Participants were initially directed to an informed consent form based on their home institution. Consenting participants then answered questions about their education level, field of study, prior course, and research experience (for undergraduates only), gender, race, and any past card-sorting activity experience before beginning the first sorting activity.

The first (unframed) sort asked participants to sort the 16 cards into groups based on "common underlying scientific principles." Participants were instructed to have no more than 15 groups and to include each card in no more than one group. Participants viewed and visually sorted the cards on a linked Flippity page, then recorded their groupings along with a group name and reasoning for the name into the survey instrument. Participants were also asked to record the time they began and ended the sorting task.

Participants then performed a second (framed) sort where they were asked to sort cards into four prenamed groups consisting of anecdotal/story, correlational/observational, experimental/manipulative, and secondary/meta-analysis. They recorded their groups, beginning and ending framed sort times, and answered a series of open-ended questions about their sorts.

Heatmap and Hierarchical Clustering

To help validate the final full sorting task, a matrix was created showing percentage pairings of each card for both undergraduate students and faculty in the unframed sort (Krieter *et al*., 2016). The undergraduate-student matrix values were then subtracted from the faculty values to generate a heat map showing which pairs were more or less frequent in faculty versus undergraduate-student sorts. The values from the faculty and undergraduate-student matrices were then analyzed for hierarchical clustering using a Ward's minimum variance method in JMP Pro.

Edit Distance and Percent Pairings Metrics

Edit distance measures the minimum number of card moves needed to transform a sort into a predicted sort and has previously been used as a measure to quantify similarity between sorts and predicted sorts (Deibel *et al.*, 2005). Three predicted sorts were used: deep-feature sort according to the research approach, surface-feature sort according to the study organism, and surface-feature sort according to the researcher. The edit distance of a sort that exactly matches the comparison-predicted sort would be zero while a sort that nearly matches the comparison sort but that misplaces three cards into other groups would be three. Thus, smaller edit distances indicate sorts more closely, matching the comparison sort. A Python program was used to calculate the edit distances of each framed and unframed participant sort to each of the three predicted deep- or surface-feature sorts.

Percent pairing is another previously used measure that quantifies the percentage of sorts which place two cards into the same group (Smith *et al.*, 2013). Thus, a high percent pairing indicates that participants frequently placed those two cards in the same group. A Python program was used to calculate the percent pairing for all possible card pairs for all participant sorts.

TABLE 1 . Demographics of participant population

		Undergraduate student	Graduate student	Faculty/postdoc
Institution	**Emory University**	527 (93%)	5 (50%)	5 (45%)
	Georgia Institute of Technology	42 (7%)	5 (50%	6 (55%)
Gender	**Female**	370 (65%)	10 (100%)	4 (36%)
	Male	189 (33%)	0 (0%)	6 (55%)
	Non-binary	4 (<1%)	0 (0%)	0 (0%)
	Other/unanswered/prefer not to say	6 (1%)	0 (0%)	1 (9%)
Ethnicity*	**African American**	64 (11%)	1 (10%)	0 (0%)
	Asian	254 (45%)	2 (20%)	1 (9%)
	Hawaiian/Pacific Islander	3 (<1%)	0 (0%)	0 (0%)
	Latinx/Chicano	74 (13%)	2 (20%)	0 (0%)
	Native American	6 (1%)	0 (0%)	0 (0%)
	White	219 (38%)	5 (50%)	8 (73%)
	Other	6 (1%)	1 (10%)	1 (9%)
	Decline to answer/Unanswered	8 (1%)	0 (0%)	1 (9%)
Position		1st year (329, 58%)	PhD candidate (8, 80%)	Faculty (10, 91%)
		2nd year (151, 27%)	Masters candidate (2, 20%)	Postdoc (1, 9%)
		3rd year (23, 4%)		
		4th year (9, 2%)		
		Unanswered (57, 10%)		
Experience/field of study		**AP Biology experience**	**Field of study**	
		AP Biology experience with a 4 or 5 on the exam (249, 44%)	Biology (4, 40%)	Biology (4, 36%)
		AP Biology experience only (105, 18%)	Public health (2, 20%)	Ecology (3, 27%)
		No AP Biology experience (211, 37%)	Biochemistry (1, 10%)	Genetics (2, 18%)
		Unanswered (4, <1%)	Neuroscience (1, 10%)	Unanswered (2, 18%)
			Cancer biology (1, 10%)	
		Research lab experience	Structural biology (1, 10%)	
		Research lab experience (126, 22%)		
		No research lab experience (439, 77%)		
		Unanswered (4, <1%)		

*Participants were included in all demographic groups they selected so may be represented in multiple groups.

Comparison of Participant-Group Sort Edit Distances

Differences between edit distances for participant groups (undergraduates, graduate students, and faculty/postdocs) and sort conditions (unframed and framed) were carried out using Kruskal-Wallis H test with post hoc Dunn's analyses.

Comparison of Demographic Information

While the faculty and graduate-student population was too small to follow up with demographic comparisons between sorting conditions, Kruskal-Wallis H tests with post hoc Dunn's analyses were carried out to compare edit distances between different demographic conditions for undergraduate sorts.

Free Response Rubric

A rubric to analyze open-ended responses to the card-sorting task was developed by identifying common themes among the responses (Table 2). Researchers utilized a subset of 42 novice responses for rubric development and training. Three members of the research team documented common themes and then collectively met to share ideas and refine rubric categories. Two other members of the research team then applied the rubric to the same subset of data and provided feedback to further refine and finalize the rubric. While we initially began this process using a grounded theory approach, we realized that upon implementation of our original coding rubric that our themes fit more accurately into a modified version that highlighted areas related to our hypothesized groupings of experts and novices with a middle category to inform responses that were more advanced than novice responses but did not fully demonstrate expert-like thinking (Strauss and Corbin, 1990). For examples of written responses and their coding scheme, please see the *Supplemental Files [card sorting cards and hypothesized sort]*.

Using the rubric, two researchers coded participant's written responses and interrater reliability was calculated to get a baseline for agreement. Cohen's Kappa was calculated for the category and subcategory level for each part of the rubric with the following values indicating moderate to substantial agreement among the researchers. For the "Grouping Explanation Data", the κ value was 0.75 at the category level and 0.51 at the subcategory level. For the "Sorting Challenges", the κ value was 0.67. For the "Preferences", the κ value was 0.58. The two researchers involved in the coding then discussed the response and coded to consensus. Responses were then grouped by whether they displayed novice-like thinking, developing thinking, expert-like thinking, or more than one category. Another author who was not involved in the rubric coding process independently coded a subsample (20) of responses to check for interrater reliability with the consensus code. Cohen's Kappa values were calculated showing substantial to near-perfect agreement among the researchers. For the "Grouping Explanation Data", the κ value was 0.82 at the category level and 0.72 at the subcategory level. For the "Sorting Challenges", the κ value was 0.87. For the "Preferences", the κ value was 0.76.

Comparing Edit Distance versus Free Response Selection

To test for internal validity of the task we compared the deep-feature edit distance to what level of thinking the responses demonstrated, as coded by the rubric. Differences in edit distance between categories of responses was analyzed by ANOVA.

TABLE 2 . Rubric for written prompt analysis

Rubric	
Category	**Subcategory**
	Reasoning
Novice-type thinking	Subject of card as group name
	Field of research
	Who conducted research
	Multiple surface features mentioned
Developing thinking	Data collection/similar data
	Similar type of conclusion
	Correlation/causation
	Control group
Expert-type thinking	Given group names similar to own group names
	Experimental design
	Lack of evidence
	Past study/meta-analysis
	Challenges
Reasoning for challenges	Did not understand anecdotal
	Card potentially in multiple groups
	Card doesn't belong in any group
	Did not understand meta-analysis
	Did not understand group names of framed sort
	Sorting Preference
Preferred framed sort	Given names clearer
	Coherence of groups
	Specificity of group names
Preferred unframed sort	Ease of understanding
	Trouble placing cards into framed sort
	Created groups were broader
No preference	Created names and given names similar
	Both sets were easy to sort
	Both sets were difficult to sort

RESULTS

Here, we sought to explore whether a card-sorting task could be used to differentiate between expert and novice thinking in scientific research. We developed 16 scenario cards designed with a deep feature (either anecdotal/story, correlational/observational, experimental/manipulative, or secondary/meta-analysis), surface-feature organism (either human, fly, plant, or microbe), and surface-feature researcher (either scientist, medical professional, student, or layperson). Participants completed both unframed (sorting 16 cards into anywhere from 2–15 groups based on "underlying scientific research principles") and framed (sorting cards into groups labeled with deep feature categories) sorts.

Expert Sorts Recreate Predicted Deep Feature Sort

To examine how novices and experts sort the scenarios in an unbiased manner we first examined the unframed sort for how often each potential card pair was sorted into the same group by participants (percent pairing metric) and percent pairing between each card were organized into a matrix for analysis. These percent pairing matrices for faculty/postdocs and undergraduates were analyzed by hierarchical clustering to provide a blind analysis to the hypothesized sorts. In this modeling, for the faculty sort, the clustering recapitulates the hypothesized

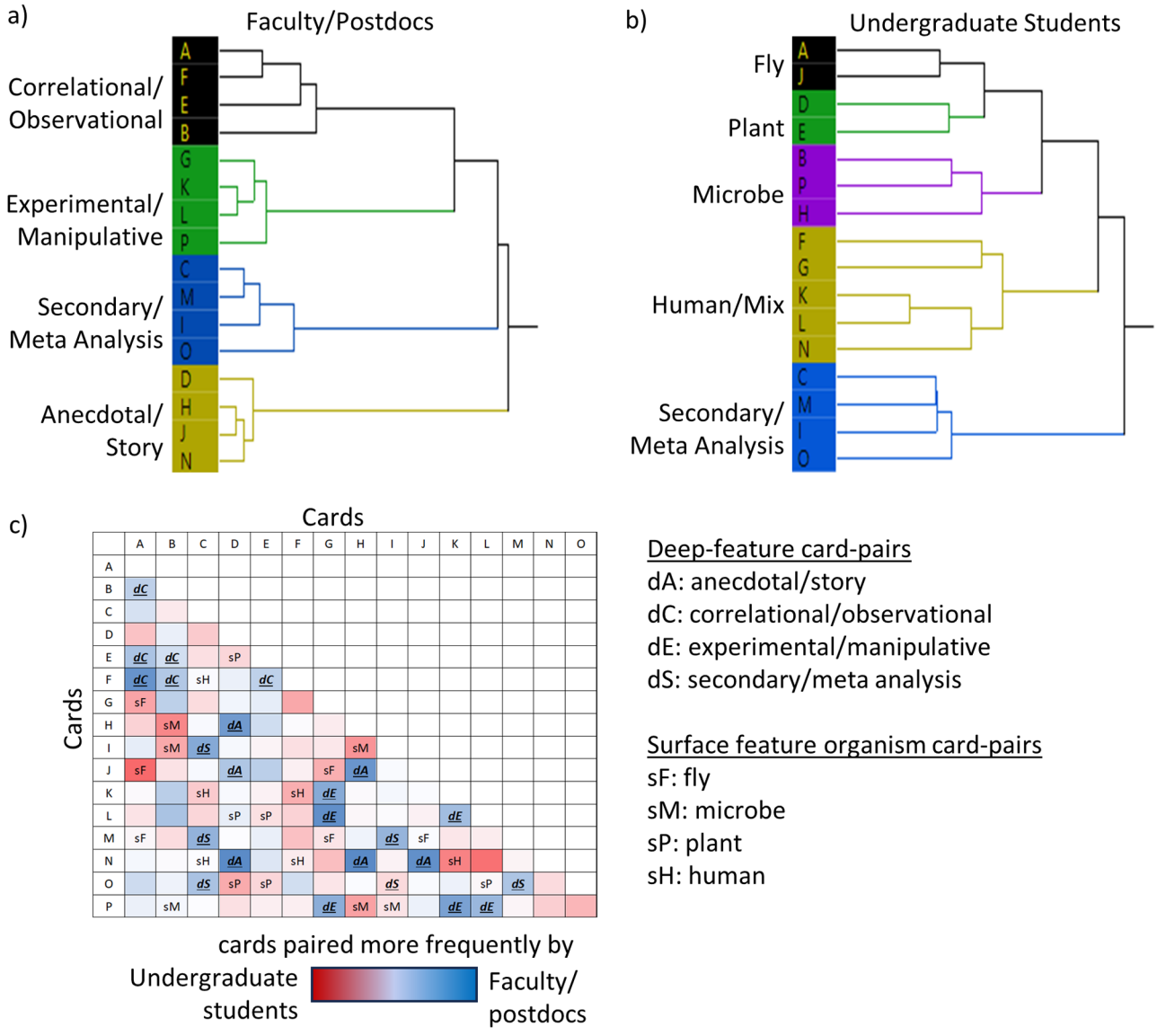

FIGURE 1. Hierarchical clustering analysis and percent pairing heatmap. Data on how many times cards in the unframed sort were paired together were used to create heat matrices of card pairings for faculty/postdoc and undergraduate students. Hierarchical clustering analysis using Ward's minimum variance method generated card groupings for (a) faculty/postdoc and (b) undergraduate students. Faculty/postdoc card groupings clustered according to predicted deep features, while undergraduate students clustered more weakly around predicted surface feature subject categories along with meta-analysis/secondary. Longer branch length indicates less tightly clustered cards. (c) Heatmap showing difference in percent pairing between faculty/postdoc and undergraduate students. The card pair for each cell in the matrix is identified by the matrix column and row card letters. The deeper blue the square the more faculty/postdocs made the pairing compared with deeper red where more undergraduates made the pairing. Letters indicate predicted deep feature pairings.

expert-like deep sort perfectly with correlational/observational, experimental/manipulative, secondary/meta-analysis, and anecdotal/story all represented as distinct clusters on the tree (Figure 1a). The undergraduate student sort largely recapitulated the predicted surface feature by organism sort, identifying a cluster for each organism type (Figure 1b). However, it did not perfectly recapitulate each category group (as not all cards cluster into their organism group) and the distances to groups are much longer than the faculty sort distances, indicating less uniformity. Additionally, the undergraduate clustering identified one deep-feature category (secondary/meta-analysis) as a cluster, although with less certainty (longer branch length) than in the faculty clustering.

To examine differences between the faculty/postdoc expert and undergraduate novice percent pairing data we created a heat map matrix of pairing differentials by subtracting the undergraduate matrix from the faculty/postdoc matrix (Figure 1c). For a majority of deep-feature pairings (23/24), experts sorted these cards more consistently together than novices. Conversely, for the majority of surface-feature organism pairings (23/24) novices sorted these cards more consistently together than experts.

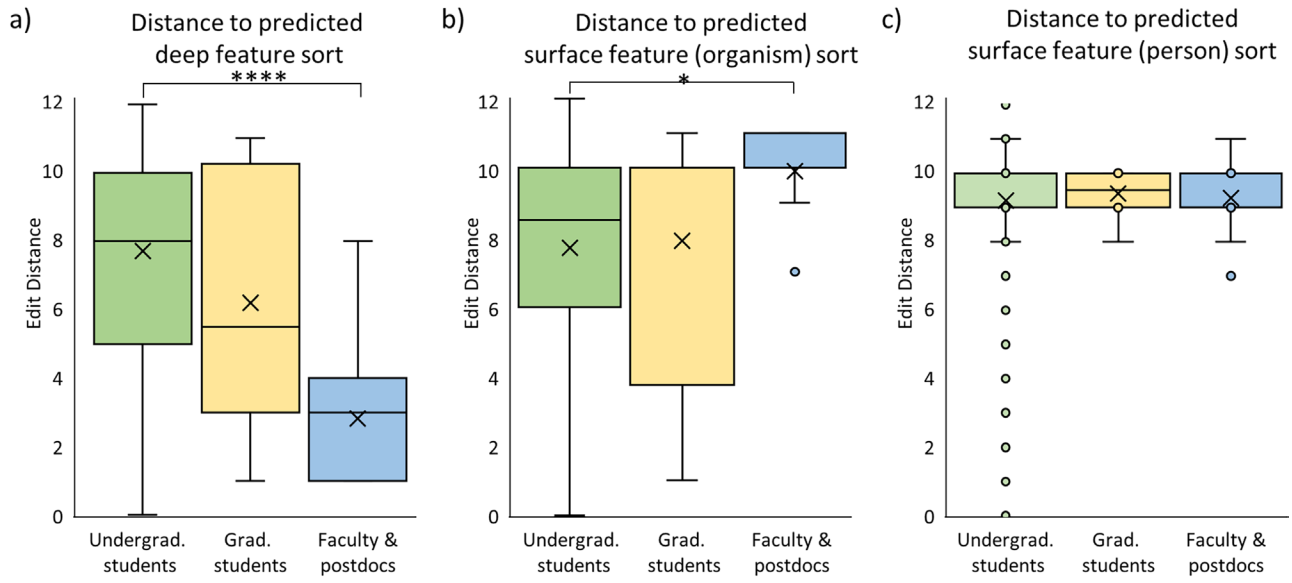

FIGURE 2. Unframed edit distances to predicted sorts for undergraduate students, graduate students, and faculty/postdocs. Box and whisker plot showing edit distances of unframed sorts to the predicted (a) deep-feature sort, (b) surface-feature organism sort, and (c) surface-feature researcher sort. Lower numbers indicate sorts more closely matching the predicted sort. Sample sizes were 569 (undergraduate students), 10 (graduate students), and 11 (faculty/postdocs). Boxes frame the middle two quartiles with an "X" for the average score and horizontal line for the median score. Whiskers indicate nonoutlier minimum and maximum scores.

Experts More than Novices Use Research Approach to Organize Cards in an Unframed Sort

In the unframed sort, participants were asked to sort based on "underlying scientific principles" but were not given further guidance as to what principles to use. We hypothesized that experts would sort more often than novices by the deep feature of research approach while novices would sort more often than experts by surface features such as who carried out the research or the subject organism used. To test this, we calculated the edit distances (minimum number of cards that would have to be moved in order to convert a sort to a predicted sort) for each participant sort relative to the three predicted sorts. Thus, edit distances closer to zero indicate sorts that closely match the comparison predicted sort while larger edit distances indicate more dissimilar sorts. Our results showed that faculty/postdocs were more likely than undergraduate students to sort based on the deep-feature of research approach (Figure 2a, Kruskal-Wallis H test for the three groups $p < 0.001$ with Post hoc Dunn's test $p < 0.0001$ comparing undergraduate and faculty/postdoc groups). Our results also show a progression from novice to expert, with undergraduate sorts being most dissimilar to the predicted deep-feature sort, followed by graduate students and then faculty/postdocs with sorts most similar to the predicted deep-feature sort (Figure 2a, average edit distances of 7.7, 6.2, and 2.8, respectively). In fact, faculty/postdoc sorts, on average, would require only three card moves to perfectly match the predicted deep-feature sort.

We also found that novices were more likely than experts to sort based on the organism featured in the scenarios as undergraduate students' edit distances to the predicted surface-feature sort based on scenario organism were significantly lower than faculty/postdocs' (Figure 2b, average edit distances of 7.7 and 9.9, respectively, Kruskal-Wallis H test for the three groups $p < 0.01$ with Post hoc Dunn's test $p < 0.01$ comparing undergraduate and faculty/postdoc groups). We did not find a difference between novices and experts in sorting scenarios based on the person featured in the scenario (Figure 2c, Kruskal-Wallis H test for the three groups $p > 0.05$).

Both Novice and Expert Sorts Became More Similar to the Predicted Deep-Feature Sort when Scaffolded with Deep-Feature Category Names

When provided with an expert-like conceptual framework by providing group names of anecdotal/story, correlational/observational, experimental/manipulative, and secondary/meta-analysis to use in the framed sorting task, we found that on average all sorts better matched the predicted deep-feature sort (Figure 3a). Average edit distances dropped from 7.7 to 5.2 for undergraduates, 6.2 to 3.2 for graduate students, and 2.8 to 1.9 for faculty/postdocs though the decrease was only statistically significant for undergraduate students (Kruskal-Wallis H test for the six groups $p < 0.001$ with Post hoc Dunn's test $p < 0.0001$ comparing undergraduate unframed and framed sorts). Conversely, we found that undergraduate student sorts became less similar to the predicted surface-feature sort based on organism in the framed sorting task (Figure 3b, average edit distance increased from 7.7 to 8.9, Kruskal-Wallis H test for the six groups $p < 0.001$ with Post hoc Dunn's test $p < 0.0001$ comparing undergraduate unframed and framed sorts). We also found a small but significant decrease in edit distance to the predicted surface sort based on person featured for undergraduate students (Figure 3c, Kruskal-Wallis H test for the six groups $p < 0.001$ with Post hoc Dunn's test $p < 0.0001$ comparing undergraduate unframed and framed sorts).

Although undergraduate-student sorts became more similar to the predicted deep-feature sort in the framed sorting task,

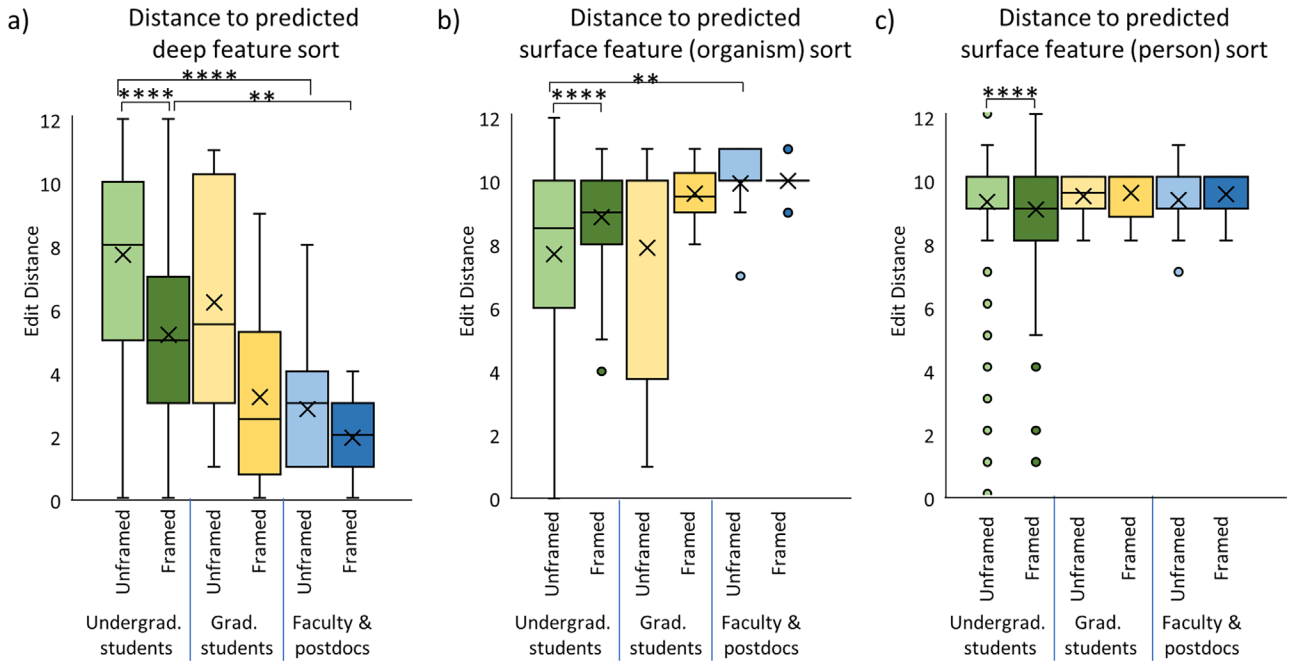

FIGURE 3. Unframed and framed edit distances to predicted sorts for undergraduate students, graduate students, and faculty/postdocs. Box and whisker plot showing edit distances of unframed and framed sorts to the predicted (a) deep-feature sort, (b) surface-feature organism sort, and (c) surface-feature researcher sort. Lower numbers indicate sorts more closely matching the predicted sort. Sample sizes were 569 (undergraduate students), 10 (graduate students), and 11 (faculty/postdocs). Boxes frame the middle two quartiles with an "X" for the average score and horizontal line for the median score. Whiskers indicate nonoutlier minimum and maximum scores.

they still did not sort as closely to the predicted deep-feature sort as faculty/postdocs (Figure 3a, Post hoc Dunn's test $p < 0.001$ comparing framed undergraduate and faculty/postdoc sorts).

Secondary/meta-analyses Cards Most Often Grouped Together by Novices

To examine whether certain deep-feature categories were more or less challenging for undergraduate students to sort as predicted, we examined the percent card pairings for these categories. We found that in the unframed sort, undergraduate students more commonly grouped secondary/meta-analysis cards together than for card pairs in other deep-feature groups (Figure 4a). We also found that undergraduate students least often grouped correlational/observational cards together compared with the other deep-feature groups. Undergraduate students grouped correlational/observational cards together about as often as they grouped cards with surface-feature groups together. We did not find large differences in the grouping of the surface-feature pairings, indicating that when organizing by surface feature, novices were not challenged by one subcategory more than others.

When given guidance on how to sort cards (i.e., given group names), undergraduate students more often grouped cards within the deep-feature groups together (Figure 4b). Interestingly, while the average percent pairings within the deep-feature groups all increased from the unframed to the framed sort for undergraduate students, they showed the biggest percent pairings increase for experimental/manipulative and secondary/meta-analysis while anecdotal/story showed a relatively small gain of around only 5%.

Undergraduates with Advanced Placement Biology Experience and Non-PEER Undergraduates Sorted More Closely to Predicted Deep-feature Sort

To examine whether past experience with Advanced Placement (AP) Biology might impact card sort measures we grouped students according to self-reported experience with AP Biology classes (experience with AP Biology and a score of at least a four on the AP Biology exam, experience with AP Biology but did not receive at least a four on the AP Biology exam, and no AP Biology experience). We found that AP Biology experience did impact students' sort similarity to the predicted deep-feature sort with students that received at least a four on the AP Biology exam having the lowest edit distances to the predicted deep-feature sort for both the unframed and framed sorts (Kruskal-Wallis H test for the six groups $p < 0.001$). The AP Biology students with at least a four on the exam had unframed sorts significantly more similar to the predicted deep-feature sort compared with students without AP Biology experience and framed sorts significantly more similar to the predicted deep-feature sort compared with students with AP Biology but without at least a four on the exam (Post hoc Dunn's test $p < 0.01$ and $p < 0.001$ respectively). All groups showed framed sorts significantly more similar to the predicted deep-feature sort than the unframed sort (Post hoc Dunn's test $p < 0.0001$ for all three comparisons).

We found that both non-PEER and PEER student sorts became more similar to the predicted deep feature sort in the framed condition to a similar extent (average decrease in edit distance was 2.6 and 2.5 respectively) but found that PEERs had significantly higher overall edit distance in the framed condition than non-PEERs (Kruskal-Wallis H test with the four

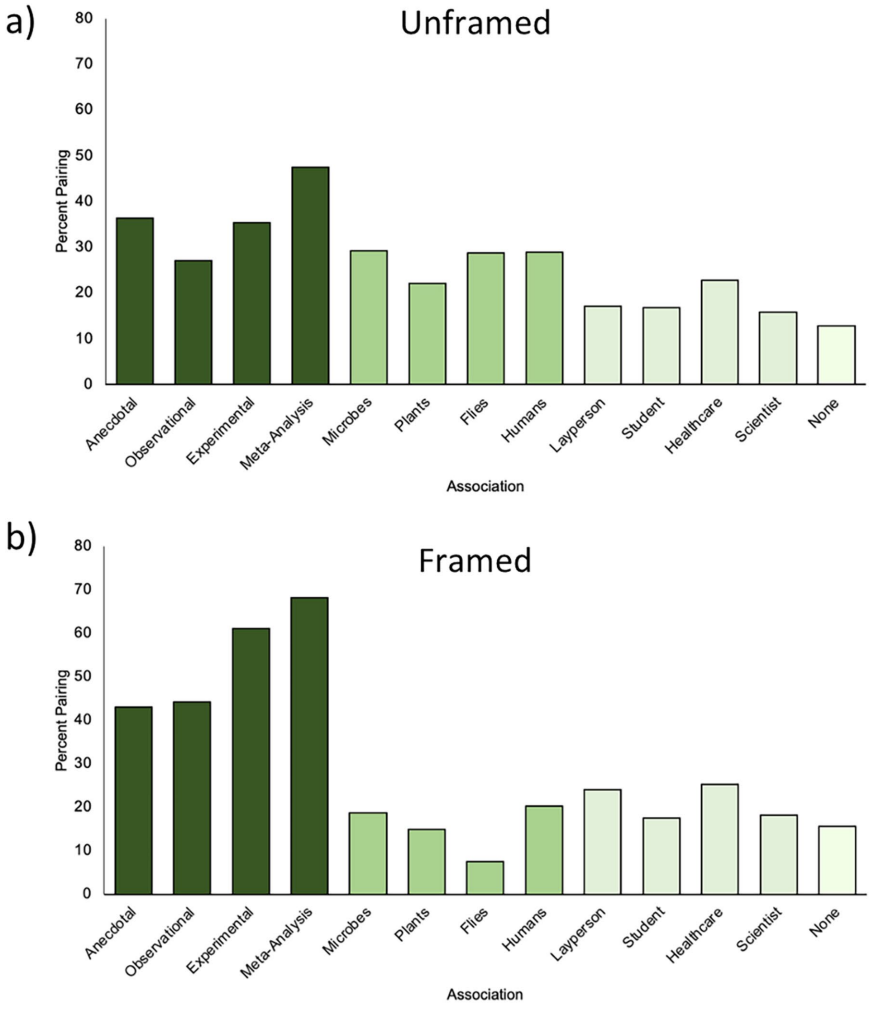

FIGURE 4. Percent predicted sort category card pairings for unframed and framed undergraduate student sorts. The average percent of sorts that grouped two cards together for each possible card pairing within predicted deep and surface feature categories are shown. Results presented for undergraduate student unframed (a) and novice framed (b) sorts. Categories colored according to their predicted sort; deep feature (dark green), surface-feature organism (medium green), and surface-feature researcher (light green).

groups $p < 0.001$ with Post hoc Dunn's test $p < 0.01$ comparing framed non-PEER and PEER groups). We also found that both female- and male-student sorts became more similar to the predicted deep-feature sort in the framed condition to a similar extent (Figure 5c, average decrease in edit distance was 2.5 and 2.7, respectively) and found no significant difference in overall edit distances between females and males for both the unframed and framed sorts (Post hoc Dunn's test $p > 0.05$ for both comparisons).

Novices Used a Mix of Novice Like, Developing, and Expert-Like Thinking While Experts Rarely Used Novice-Like Thinking

We used a rubric to code written responses into three categories (expert-like thinking, novice-like thinking, and developing thinking, a transitional category between expert and novice thinking) with four subcategories each (Table 2). Novice-like thinking was used by 40% of undergraduate students with 32% mentioning using the "subject of the card as the group name" (Figure 6a). The majority of undergraduate students expressed developing thinking with 51% having responses fall into this category. Of those, the overwhelming majority discussed "data collection/type of data" in their response with 42% of all undergraduate students using this framing (Figure 6a). Expert-like thinking was expressed by 32% of undergraduate students with "past study/meta-analysis" leading the way at 22%. Of the undergraduate-student responses, 10% expressed both novice-like and developing thinking, 30% expressed expert-like and developing thinking, 5% expressed both novice- and expert-like thinking.

In contrast, faculty/postdocs used expert-like reasoning in their written responses more than the other two categories with 100% of faculty/postdocs displaying expert-like thinking. "Given group names similar to own group names" (90%), "past study/meta-analysis" (90%), and "experimental design" (82%) were represented in the vast majority of faculty (Figure 6c). Developing thinking was represented 36% of the time with "similar type of conclusion" and "data collection method" mentioned 27% and 18% of the time, respectively (Figure 6c). Each time an expert participant used developing thinking they also displayed expert-like thinking. Novice-like thinking was never displayed by the faculty/postdoc group. The majority of faculty/postdocs (7/11, 64%) displayed only expert-like thinking while some displayed both expert and developing thinking (4/11, 36%). Graduate students were mixed with 60% displaying expert-like thinking, 40% displaying developing thinking, and 30% displaying novice-like thinking (one participant had an uncodeable response). Of the expert-like reasoning displayed by graduate students, "given group names", "experimental design", and "past-study/meta-analysis" were all present in all the graduate students who displayed this type of thinking (60%; Figure 6b). All the developing responses were of the "data collection/similar data" subcategory and the novice responses were "field of research" and "subject of card as group name" (20%).

Card Potentially Being in Multiple Groups was the Most Consistent Challenge

We also asked participants what challenges they had sorting cards during the second framed sort and analyzed this data for undergraduate students. Five types of reasoning were present

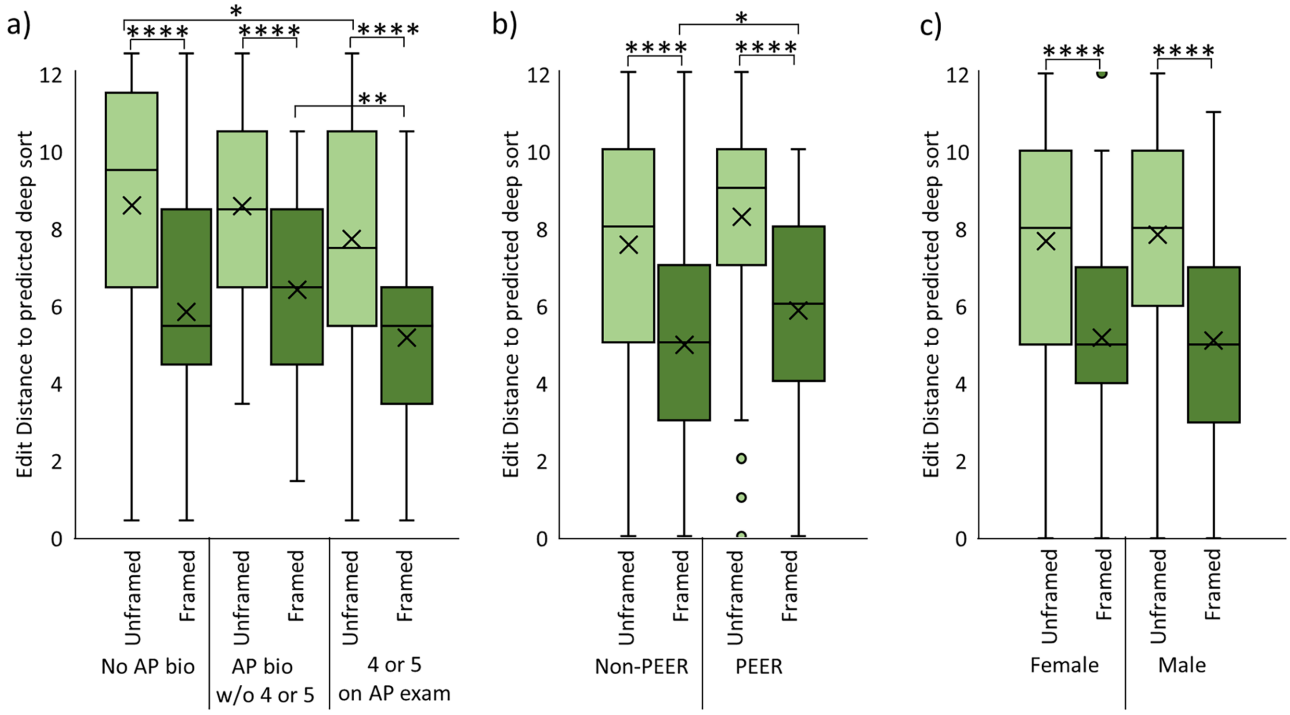

FIGURE 5. Unframed and framed edit distances to predicted deep-feature sort for different undergraduate groups. Box and whisker plot showing edit distances of unframed and framed sorts to the predicted deep-feature sort for (a) undergraduate students with no high-school AP Biology experience ($n = 211$), students with high-school AP Biology experience but without a high exam score ($n = 105$), and students with high-school AP Biology experience and a high exam score ($n = 249$), (b) non-PEER ($n = 423$) and PEER ($n = 139$) students, and (c) female ($n = 370$) and male ($n = 189$) students. Lower numbers indicate sorts more closely matching the predicted deep-feature sort. Boxes frame the middle two quartiles with an "X" for the average score and horizontal line for the median score. Whiskers indicate nonoutlier minimum and maximum scores.

consistently across our undergraduate-student responses. 30% of undergraduate students indicated that a card potentially being in multiple groups was a challenge. 12% of undergraduate students indicated a lack of understanding of the anecdotal/story category and 7% indicated a lack of understanding of the meta-analysis category. Finally, 6% discussed a card not belonging in any group, and 3% did not understand group names of the framed sort (figure 6d).

Undergraduate Students Preferred the Framed Sort to the Unframed Sort

Undergraduate student participants overall preferred the framed sort over the unframed sort with 51% of participants preferring the framed sort and only 16% preferring the unframed sort (Figure 5d). Three reasons for preferring the framed sort were coded in rubric analysis with "given names clearer" (29%) being the most common, followed by "coherence of groups" (13%), and specificity of group names (10%; Figure 6e). For the 16% of novice participants who preferred the unframed sort, the top reasoning given was "ease of understanding" (11%). 14% of participants had no preference with the top reason cited as "both were easy to sort" (9%).

Rubric Coding Correlated with Edit Distance

When we examined how rubric coding related to edit distance for all participants to the predicted deep feature sort we found that individuals whose rubric coding was "novice only" had the highest average edit distance (9.14, $n = 184$), followed by in order: "novice/developing" (7.66, $n = 38$), "developing only" (6.14, $n = 91$), "developing/expert" (4.90, $n = 172$), and "expert" (3.76, $n = 25$; Figure 7). The difference in edit distance between these different groupings was significant (Kruskal-Wallis H test for the five groups $p < 0.001$) and all pairwise comparisons between nonadjacent groups were significant (Post hoc Dunn's test $p < 0.01$ in all cases). Individuals who displayed all three types of thinking in their code ($n = 24$) had an edit distance 6.16 showing high similarity to the average for those coding as developing only while individuals who coded as both novice and expert were rare ($n = 8$) and had an edit distance of 5.13, between "developing" and "developing/expert."

DISCUSSION

With the increasing use of course-based research experiences in colleges and universities, it is important to examine how these experiences impact students' development of scientific research skills. Ideally, research experiences would help move students from novice-like to more expert-like thinking in how they conceptualize research. Card-sorting tasks where participants are asked to group cards with similar problems or scenarios together have been used previously to measure differences in how experts and novices organize materials. Novices tend to organize cards based on surface-level similarities while experts tend to organize cards based more on underlying conceptual similarities. However, card sort tasks have not previously been

FIGURE 6. Rubric analysis of written prompt responses. Undergraduate student (a), graduate student (b), and faculty/postdoc (c) sorting reasonings organized by expert-like thinking (blue), developing thinking (yellow), novice-like thinking (green). For (a), (b), and (c) percentages sum to gt 100 as subjects could be categorized as using multiple methods of reasoning. Reasons given by novices for sorting challenges (d). Undergraduate student card-sort preferences (e) and reasons for card-sort preferences (f) with no preference (yellow), preferred unframed sort (green), and preferred framed sort (blue). For (c), (d), and (e) percentages total to lt 100 as not all subjects responded to prompts.

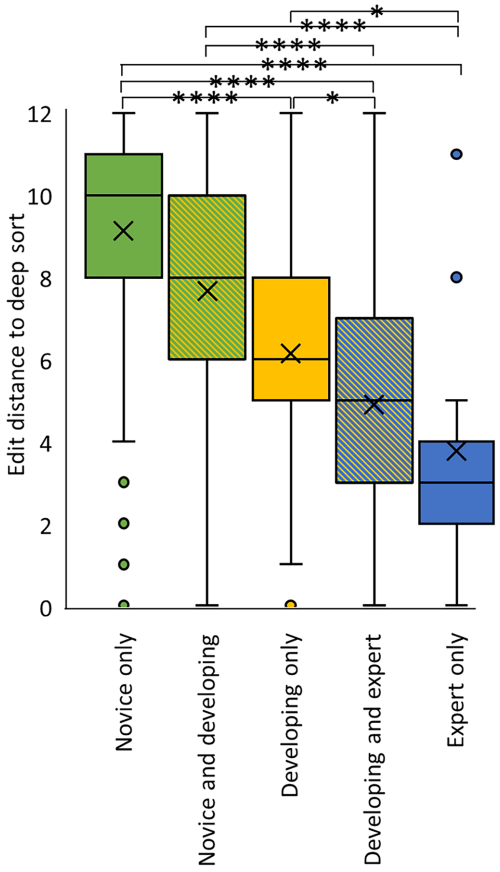

FIGURE 7. Deep-feature edit distance based on rubric coding. Rubric coded responses were grouped together based on what type of sorting reasoning they displayed and then average deep-feature edit distances for each group was compared. Shown is a box and whisker plot comparing respondents who utilized novice thinking only ($n = 184$), novice and developing thinking ($n = 38$), developing thinking only ($n = 91$), developing and expert thinking ($n = 172$), and expert thinking only ($n = 25$). Lower edit distance indicates sort more closely matching the predicted deep-feature sort. Boxes frame the middle two quartiles with an "X" for the average score and horizontal line for the median score. Whiskers indicate nonoutlier minimum and maximum scores.

used to measure expertise in research. Here, we developed and validated a card-sorting task aimed at measuring novice versus expert thinking in scientific research.

Our results show that we can detect a measurable difference between experts and novices when they are simply asked to group similar scenario cards together. Using independent hierarchical clustering analysis we were able to demonstrate that our sorting task was designed in a manner that was consistent with our hypothesized surface and deep-feature sorts. Our results supported our hypothesis that experts would be more likely to group cards according to fundamental differences in the research approach described on cards than novices. This suggests that experienced researchers may find how research is performed to be of high importance as they know that the mode of research impacts the types of questions which can be addressed and the conclusions that can be drawn (e.g., manipulative experiments can address questions of causation while observational studies can only address questions of correlation). Research experts were also less likely than novices to group cards according to the organism used in a study. Experts are likely more familiar with connecting findings from numerous studies using different model organisms to build understanding of biological processes. Both experts and novices grouped cards least often according to the person carrying out the research, perhaps recognizing that this does not impact the conclusions that can be drawn from a study, or due to instructions provided for the initial sort about looking for the "underlying scientific principles" priming them away from this framing option.

When participants were provided expert-like scaffolding via group names based on research approach and then asked to sort cards into those groups (framed sort), both novices and experts were better able to sort into those categories. On average, novices who were given category titles were able to sort cards into those categories about as well as or better than graduate students who were not given category names. This suggests that novices are largely able to recognize different research approaches when instructed to focus on that but that they were less likely than experts to view the research approach as the most important distinction between studies in their initial unframed groupings. Thus, the shift from novice to expert-level thinking in biological research may, in part, involve placing more emphasis on experimental design over other details such as model organism used.

Looking at which cards novices grouped together more or less frequently (percent pairings) for our novices gives us some additional insight into areas where they may have struggled with the task. Of the deep-feature sorting groups, students seemed to have the easiest time identifying meta-analysis type studies. This could be because the meta-analysis studies were more obvious in their approach, or perhaps the way in which these novices have been exposed to science previously primed them to be able to identify these types of studies. Interestingly, anecdotal/story and observational/correlational have a much smaller increase between unframed and framed in percent pairings compared with experimental/manipulative and secondary/meta-analysis. This may indicate that students are having trouble identifying these two categories, so directed instruction on these research approaches may be a fruitful area to explore for developing research expertise in undergraduates.

When we compared undergraduates with varying levels of past biology curriculum experience as measured by exposure to high school AP Biology courses and performance on the AP Biology exam, we found that undergraduates with more high-school-level biology have more expert-like thinking on our card-sort task. This may indicate that people who have taken and excelled in AP Biology courses have developed more expert-like research thinking. These results are consistent with findings from other studies about performance on the AP exam and college performance, but it would be interesting to identify exactly what type of thinking is being promoted by these exams or if it is a selection bias issue (Sadler and Tai, 2007; Ackerman *et al.*, 2013; Beard *et al.*, 2019).

In comparing PEER and non-PEER students, we found that non-PEER students exhibited more expert-like thinking on our sort task than PEER students. This is consistent with literature related to preparedness for college and suggests that interventions in this area could be beneficial to students from PEER

groups when preparing for lab courses (Stephens *et al.*, 2014; Estrada *et al.*, 2016; Broda *et al.*, 2018). Our card-sort task did not measure a difference between female and male students.

When we examined participants' written responses on their thought process behind their unframed sorts we found evidence of novice-like, expert-like, and developing thinking. Subjects who expressed novice-like thinking were just focused on the surface features of the cards and not connecting to the deeper research and experimental-design components. Conversely, subjects expressing expert-like thinking were recognizing the deeper research connections between the different cards experimentally. Those in the middle who expressed developing thinking understood that there was some deeper meaning to the cards but may have been unable to explicitly connect with those deeper principles. While students in the developing thinking demonstrated understanding of some features of experimental design, these features are more about the experimental set up and structure rather than a deeper aspects of research related to the underlying principles of the experiment and how that informs our understanding of the data and conclusions. Subjects in this category may have also expressed expert or novice-like thinking with their developing thinking. While the expert responses were useful for comparison purposes, a deeper examination of the novice responses and characterizations by the rubric can be useful for the way in which we can approach our students' perceptions and understanding of research and experiments.

We found novices' reasoning to range across novice-like, developing, and expert-like, with 40% of novices expressing novice-like thinking, 51% expressing developing thinking, and 34% expressing expert-like thinking. These reasoning groupings correlated to the deep feature sort with individuals displaying novice-only thinking being the farthest from the deep-feature sort, followed by individuals with a mix of novice and developing thinking, then developing-only, developing/expert, and expert-only thinking. Novices that express novice-like thinking may initially have greater struggles in understanding the how and why of their experiments when taking inquiry-based and CURE lab classes. Students at this level may benefit from shifting their thinking to the developing level rather than try to have them understand at an expert-like level. Getting these students to think about what type of data they are collecting and how they are collecting it could help them to move their conceptual thinking to the next level. In addition, having a holistic approach to the data and experiment could help students avoid overfocusing on the surface level details.

Once students are in the developing category of thinking, they are clearly thinking about experimental design on another level and have mostly moved past the basic surface-level structure. However, students in this category may end up too focused on the details of the data and what was done in the experiment as opposed to the larger implications behind the conclusions. There is evidence that students can observe patterns in data, but it is more difficult to get them to understand conclusions based on that analysis (Germann and Aram, 1996; Cary *et al.*, 2019). This can be especially apparent when students are tasked with designing their own experiments in CUREs. In a well-designed experiment, scientists should think ahead to what their potential results may look like and how they will analyze those results, based on how they have set up their experiment. Students that are at this level of thinking may be structuring their ideas into the basics of how to collect the data rather than what the data will look like and how that will influence the conclusions they can draw. This shows up in aspects of experimental design and research related to thinking about all the different variables that can influence an experiment and how best to control for them. By carefully designing activities in CUREs and other labs, we can potentially shift this group's thinking into more expert-like concepts focused on the type of experiment performed and why that is important.

Students in the expert-like thinking category need opportunities to hone their thinking and utilize it in a practical manner. Students who are already showing expert-like thinking are in position to gain next-level scientific processing skills and can greatly benefit from the opportunities that CUREs and inquiry-based labs provide. They may also be a valuable resource for helping other students to understand the concepts and conclusions of the work in group settings as they may have a good understanding of where other students are struggling. In all cases, students preferred the frame sort over the unframed sort, indicating that students at all levels can benefit from discussing the importance of research approach openly.

One potential limitation for this study is that roughly 20% of the student population had some kind of difficulty with their understanding of the different category titles in the framed sort. This may indicate that some of the sorting difficulty could be due to unfamiliarity with the category titles, although an argument could be made that if a student does not understand a category title, they may not have a good understanding of the underlying principle. Another potential difficulty could be the use of the word "observes" on two cards not in the "correlational/observational" category that could lead to miscuing for students. Even with these difficulties, 51% of students preferred the framed sort. The large shift in edit distance toward a more deep-feature understanding in the framed sort compared with the framed sort indicates that some scaffolding to student thinking is useful at this point in their education process.

Another potential limitation of this study is that our "unframed" sort still prompted students to sort based on "common underlying scientific principles" so may have primed them to focus more on the deep feature. It is possible that without this guidance we would find larger differentials between novices and experts or, possibly, find some experts sorting based on surface features.

Ultimately, our study has shown that there is an appreciable difference in the ways novices and experts organize their information around research. Our results suggest some ways in which laboratory science education can improve outcomes for students as we try and develop a more expert-like way of thinking about science and research. While we have demonstrated the viability of using a card-sort system for exploring differences in thinking related to research, it is unclear whether the test could be used for more diagnostic purposes to see how much progress students have made while taking biology lab courses. Ideally, we would be able to use this system to test individual-student progress over the course of a term or year to see how students have grown and adjust course curricula as needed.

REFERENCES

Ackerman, P. L., Kanfer, R., & Calderwood, C. (2013). High school advanced placement and student performance in college: STEM majors, non-STEM majors, and gender differences. *Teachers College Record*, *115*(10), 1–43.

Bangera, G., & Brownell, S. E. (2014). Course-based undergraduate research experiences can make scientific research more inclusive. *CBE—Life Sciences Education*, *13*(4), 602–606.

Beard, J.J., Hsu, J., Ewing, M., & Godfrey, K. E. (2019). Studying the relationships between the number of APs, AP performance, and college outcomes. *Educational Measurement Issues and Practice*, *38*(4), 42–54.

Bedard, J., & Chi, M. T. H. (1991). Expertise. *Current Directions in Psychological Science*, *1*(4), 135–139.

Bissonnette, S. A., Combs, E. D., Nagami, P. H., Byers, V., Fernandez, J., Le, D., ... & Tanner, K. D. (2017). Using the biology card sorting task to measure changes in conceptual expertise during postsecondary biology education. *CBE—Life Sciences Education*, *16*(1), ar:14.

Broda, M., Yun, J., Schneider, B., Yeager, D. S., Walton, G. M., & Diemer, M. (2018). Reducing inequality in academic success for incoming college students: A randomized trial of growth mindset and belonging interventions. *Journal of Research on Educational Effectiveness*, *11*(3), 317–338.

Cary, T. L., Wienhold, C. J., & Branchaw, J. (2019). A biology core concept instrument (BCCI) to teach and assess student conceptual understanding. *CBE—Life Sciences Education*, *18*(3), ar:46.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121–152.

Chi, M. T. H., Glaser, R., & Farr, M. J. (ed.). (1988). *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In *Advances in the psychology of human intelligence*, ed. R. J. Sternberg, Vol. *1*, Hillsdale, NJ: Lawrence Erlbaum Associates, 7–76.

Corwin Auchincloss, L., Laursen, S. L., Branchaw, J. L., Eagan, K., Graham, M., Hanauer, D. I., ... & Dolan, E. L. (2014). Assessment of course-based undergraduate research experiences: A meeting report. *CBE—Life Sciences Education*, *13*, 29–40.

Creamer, E. G., Magolda, M. B., & Yue, J. (2010). Preliminary evidence of the reliability and validity of a quantitative measure of self-authorship. *Journal of College Student Development*, *51*, 550–562.

Deane, T., Nomme, K., Jeffery, E., Pollock, C., & Birol, G. (2014). Development of the biological experimental design concept inventory (BEDCI). *CBE—Life Sciences Education*, *13*, 540–551..

Deibel, K., Anderson, R. J., & Anderson, R. E. (2005). Using edit distance to analyze card sorts. *Expert Syst*, *22*, 129–138.

Dreyfus, S. E. (2004). The 5-stage model of adult skill acquisition. *Bulletin of Science, Technology, & Society*, *24*(3), 177–181.

Dreyfus, H. L., & Dreyfus, S. E. (2005). Peripheral vision: Expertise in real world contexts. *Organization Studies*, *26*(5), 779–792.

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the short grit scale (GRIT–S). *Journal of Personality Assessment*, *91*, 166–174.

Ericsson, K. A., & Smith, J. (1991). Prospects and limits of the empirical study of expertise: An introduction. In Ericsson, K. A., & Smith, J. (Eds.), *Towards a General Theory of Expertise: Prospects and Limits* (pp. 1–38). New York, NY: Cambridge University Press..

Estrada, M., Burnett, M., Campell, A. G., Campbell, P. B., Denetclaw, W. F., Gutierrez, C. G., ... & Zavala, M. (2016). Improving Underrepresented Minority Student Persistence in STEM. *CBE—Life Sciences Education*, *15*(3), es:5.

Facione, P. A. (1991). Using the California critical thinking skills test in research, evaluation and assessment. La Cruz Avenue, Millbrae, CA: California Academic Press. Retrieved November 16, 2022, from https://www.insightassessment.com/product/cctst#sthash.0cnhglYK.dpbs

Galloway, K. R., Wah Leung, M., & Flynn, A.B. (2018). A comparison of how undergraduates, graduate students, and professors organize organic chemistry reactions. *Journal of Chemical Education*, *95*(3), 355–365.3.

Germann, P. J., & Aram, R. J. (1996). Student performances on the science processes of recording data, analyzing data, drawing conclusions, and providing evidence. *Journal of Research in Science Teaching*, *33*(7), 773–798.

Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguements. *CBE—Life Sciences Education*, *11*, 364–377.

Graham, M. J., Frederick, J., Byars-Winston, A., Hunter, A. B., & Handlesman, J. (2013). Increasing persistence of college students in STEM. *Science*, *341*(6153), 1455–1456.

Hanauer, D. I., Graham, M. J., & Hatfull, G. F. (2016). A measure of college student persistence in the sciences (PITS). *CBE—Life Sciences Education*, *15*(4), ar54

Harrison, M., Dunbar, D., Ratmansky, L., Boyd, K., & Lopatto, D. (2011). Classroom-based science research at the introductory level: Changes in career choices and attitude. *CBE—Life Sciences Education*, *10*(3), 279–286.

Hernandez, P. R., Agocha, V. B., Carney, L. M., Estrada, M., Lee, S. Y., Loomis, D., ... & Park, C. L. (2020). Testing models of reciprocal relations between social influence and integration in STEM across the college years. *PLoS One*, *15*(9)

Hoskinson, A., Middlemis Maher, J., Bekkering, C., & Ebert-May, D. (2017). A problem-sorting task detects changes in undergraduate biological expertise over a single semester. *CBE—Life Sciences Education*, *16*(2), ar21.

Irby, S. M., Phu, A. L., Borda, E. J., Haskell, T. R., Steed, N., & Meyer, Z. (2016). Use of a card sort task to assess students' ability to coordinate three levels of representation in chemistry. *Chemistry Education Research Practice*, *17*(2), 337–352.

Krieter, F. E., Julius, R. W., Tanner, K. D., Bush, S. D., & Scott, G. E. (2016). Thinking like a chemist: Development of a chemistry card-sorting task to probe conceptual expertise. *Journal of Chemical Education*, *93*(5), 811–820.

Kuh, G. D. (2008). High-impact educational practices: What they are, who has access to them, and why they matter. *Association for American Colleges and Universities*, Washington, DC:.

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. S. (2002). Views of nature of science questionnaire: Toward valid and meaningful assessment of learners' conceptions of nature of science. *J. Res. Sci. Teach*, *39*, 497–521.

Lin, S.-Y., & Singh, C. (2010). Categorization of quantum mechanics problems by professors and students. *European Journal of Physics*, *31*(1), 57–68.

Lopatto, D., Hauser, C., Jones, C. J., Paetkau, D., Chandrasekaran, V., Dunbar, D., ... & Elgin, S. C. R. (2014). A central support system can facilitate implementation and sustainability of a classroom-based undergraduate research experience (CURE) in Genomics. *CBE—Life Sciences Education*, *13*(4), 711–723

Makarevitch, I., Frechette, C., & Wiatros, N. (2015). Authentic research experience and "big data" analysis in the classroom: Maize response to abiotic stress. *CBE—Life Sciences Education*, *14*(3), ar27.

Mason, A., & Singh, C. (2011). Assessing expertise in introductory physics using categorization task. *Physical Review Special Topics - Physics Education Research*, *7*(2), 1–17.

Robnett, R. D., Chemers, M. M., & Zurbriggen, E. L. (2015). Longitudinal associations among undergraduates' research experience, self-efficacy, and identity. *Journal of Research in Science Teaching*, *52*(6), 847–867.

Rodenbusch, S. E., Hernandez, P. R., Simmons, S. L., & Dolan, E. L. (2016). Early engagement in course-based research increases graduation rates and completion of science, engineering, and mathematics degrees. *CBE—Life Sciences Education*, *15*(2), ar20.

Russell, S. H., Hancock, M. P., & McCullough, J. (2007). Benefits of undergraduate research experiences. *Science*, *316*(5824), 548–549.

Sadler, P. M., & Tai, R. H. (2007). Advanced placement exam scores as a predictor of performance in introductory college biology, chemistry and physics courses. *Science Educator*, *16*(2), 1–19.

Seymour, E., Hunter, A. B., Laursen, S. L., & Deantoni, T. (2004). Establishing the benefits of research experiences for undergraduates in the sciences: First findings from a three-year study. *Science Education*, *88*, 493–534.

Shortlidge, E. E., Bangera, G., & Brownell, S. E. (2016). Faculty perspectives on developing and teaching course-based undergraduate research experiences. *BioScience*, *66*(1), 54–62.

Sirum, K., & Humburg, J. (2011). The experimental design ability test (EDAT). *Bio J Col Biol Teach*, *37*(1), 8–16.

Smith, J. I., Combs, E. D., Nagami, P. H., Alto, V. M., Goh, H. G., Gourdet, M. A. A., ... & Tanner, K. D. (2013). Development of the biology card sorting task to measure conceptual expertise in biology. *CBE—Life Sciences Education*, *12*(4), 628–644.

Smith, M. U. (1992). Expertise and the organization of knowledge: Unexpected differences among genetic counselors, faculty, and students on problem categorization tasks. *Journal of Research in Science Teaching*, *29*(2), 179–205.

Stephens, N. M., Hamedani, M. G., & Destin, M. (2014). Closing the social-class achievement gap: A difference-education intervention improves first-generation students' academic performance and all students' college transition. *Psychological Science*, *25*(4), 943–953.

Stiggins, R. J. (2004). *Classroom assessment for student learning: Doing it right—using it well*. Portland, OR: Assessment Training Institute:

Strauss, A., & Corbin, J. M. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage Publications.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*, 257–285.

Zelaya, A. J., Blumer, L. S., & Beck, C. W. (2022). Comparison of Published Assessments of Biological Experimentation as Mapped to the ACE-Bio Competence Areas. In Pelaez, N. J., Gardner, S. M., & Anderson, T. R. (Eds.), *Trends in teaching experimentation in the life sciences. Contributions from biology education research*. New York, NY: Springer, Cham. https://doi.org/10.1007/978-3-030-98592-9_14