JOURNAL OF CHEMICAL EDUCATION

Article

pubs.acs.org/jchemeduc

Thinking Like a Chemist: Development of a Chemistry Card-Sorting Task To Probe Conceptual Expertise

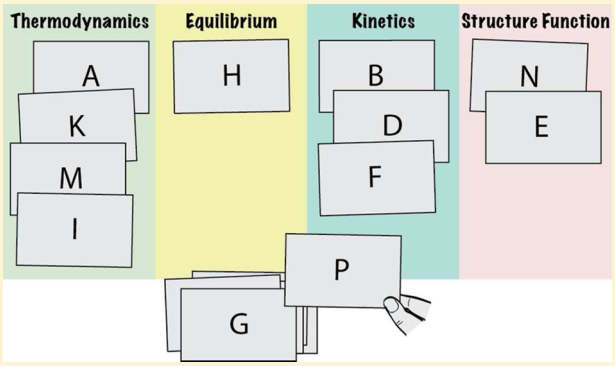
Felicia E. Krieter,[†] Ryan W. Julius,[†] Kimberly D. Tanner,[‡] Seth D. Bush,[†] and Gregory E. Scott*[,†]

[†]Department of Chemistry and Biochemistry, California Polytechnic State University, San Luis Obispo, California 93407, United States

[‡]Department of Biology, San Francisco State University, San Francisco, California 94132, United States

Supporting Information

ABSTRACT: An underlying goal in most chemistry curricula is to enable students to think like chemists, yet there is much evidence to suggest that students can learn to solve problems without thinking conceptually like a chemist. There are few tools, however, that assess whether students are learning to think like Ph.D. faculty, putative experts in the field. Here, we present a card-sorting task that probes how individuals organize information about problems in chemistry. Chemistry faculty tend to organize around "deep" features centered on fundamental ideas in chemistry while novices tend to organize around "surface" features such as problem presentation or specific vocabulary. We used established statistical techniques from card-sorting tasks in other fields and introduce a new quantitative measure that compares individual performance on the sorting task to faculty and novices that is hypothesis-independent. Initial results indicate that the card-sorting task is effective at distinguishing between populations of faculty and novices in chemistry and can be used to track progress toward more expert-like thinking over time through a chemistry education program.



KEYWORDS: *First-Year Undergraduate/General, Upper-Division Undergraduate, Chemical Education Research, Assessment, Learning Theories*

FEATURE: Chemical Education Research

A goal of many chemistry curricula is to train students to think like chemists. There is, however, much evidence that students can learn to solve specific sets of problems in chemistry without developing a fundamental understanding of the underlying concepts.[1,2] This lack of conceptual understanding and the associated alternative conceptions often persist through and beyond undergraduate training.[3] Despite this goal of preparing chemistry students to think conceptually like experts in the field, there are few assessment tools to probe this type of growth.

Some of the existing tools for examining conceptual knowledge in chemistry include the Chemistry Concept Inventory,[4] the ChemQuery[5] system, and concept mapping.[6,7] Concept Inventories are effective in investigating alternate conceptions but do not ask the question of whether students are developing an ability to organize information like an expert in the field. More open-ended approaches, like interviews and the ChemQuery system, give a richer picture of conceptual connections, but are relatively complicated to implement. A categorization task is a complementary tool to instruments like concept inventories where the emphasis is on measuring organization of content knowledge.

In cognitive psychology, expertise in a field implies not only a larger body of knowledge but also that the information is better organized and allows an individual to perform better on domain-specific tasks.[8] Many researchers are interested in how to quantify the complexity of sophistication present in the chemical thinking of experts as well as in developing frameworks for the development of expert-like thinking. As Stains and Talanquer have pointed out, expertise is not necessarily developed in a linear fashion with academic training.[9] Consequently, Sevian and Talanquer have suggested that learning progressions in chemistry should not only develop disciplinary knowledge and skills, but also focus on cross-cutting disciplinary concepts and assessment of conceptual sophistication.[10] The Perspectives of Chemists Framework also articulates that scientific reasoning with domain knowledge should be emphasized.[5] All of this suggests that conceptual expertise requires sophisticated reasoning and that assessments that focus on isolated concepts are insufficient for assessing expert-like thinking in chemistry.

Although measuring the ability of individuals to answer questions or solve problems about specific concepts may

Received: December 10, 2015
Revised: February 25, 2016

DOI: 10.1021/acs.jchemed.5b00992
J. Chem. Educ. XXXX, XXX, XXX−XXX

require better organization of domain-specific knowledge, these types of tasks do not directly measure that organization. Categorization tasks can provide insight into the knowledge organization component of expertise. A seminal categorization task in physics education by Chi et al. asked two groups with different levels of training in physics to sort physics problems based on similarity of solution.[11] The group of doctoral students, which the authors termed experts, tended to sort the problems based on underlying conceptual features in physics, or "deep features." A group of undergraduates with limited training in physics, termed novices, tended to sort by more superficial features related to the presentation of the problem, or "surface features." The use of surface and deep feature theoretical frameworks by novices and experts has been identified in other fields in the physical sciences.[9,12]

The approach taken by Chi et al. has appeared in a variety of card-sorting tasks used in psychology and has been developed for a number of different applications in other fields.[13] In addition to the work described above, there have been several card-sorting tasks developed for use in physics education research.[14−17] In biology, a card sorting task has been developed that uses hypothesized conceptual frameworks for novices and experts to measure conceptual expertise.[18] In chemistry, sorting tasks have been used to examine how novice and expert populations organize items from chemistry, emphasizing differences between multimedia and chemistry-based theoretical frameworks.[19] Our approach builds on this but differs in some key ways: our hypothesized deep and surface feature theoretical frameworks are both tied to chemistry content, we use and introduce several quantitative methods for examining differences in populations, and we have a substantially larger population of participants.

Building on the methodology of Smith et al.,[18] we have developed a card-sorting task for use in chemistry to measure differences in conceptual expertise in populations with varying degrees of chemistry training. Here, we introduce a set of methods for using a card-sorting task for measuring conceptual expertise in chemistry. Specifically, we aim to address the following research questions:

1. Can a simple card sorting exercise distinguish between populations of novices and putative experts in chemistry?
2. Is there an effective way to measure how expert-like an individual card-sorting result is without measuring against an experimenter's expectation of how experts will organize information?
3. What can the sorting task tell us about the development of expert-like thinking through a chemistry education curriculum as a tool for program assessment?

METHODS

Building the Cards

Following the model described by Smith et al.,[18] we designed a deck of 16 chemistry question cards to use as the basis of our sorting activity. We identified four categories of hypothesized surface chemistry features (reactions, acid/base, molecules, energy) and four categories of hypothesized deep chemistry features (structure/function, kinetics, thermodynamics, equilibrium). Figure 1A shows an example that was designed with the surface feature "Reactions" and the deep feature "Equilibrium".

The set of cards was designed such that each card contained one hypothesized surface feature and one hypothesized deep feature (see Figure 1B). The questions were open-ended,

A. Sample Question Card "L"

Consider the formation reaction for iron complex:

$$Fe^{3+}(aq) + 6CN^{-}(aq) \rightleftharpoons [Fe(CN)_6]^{3-}(aq)$$

At 25 °C, the formation constant for this complex is 4×10^{43}. Would adding $Fe^{3+}(aq)$ to water contaminated with cyanide (CN^-) be an effective means for removing free CN^- (aq) from solution? Explain.

L

B. Hypothesized Sorts

Hypothesized Surface Features	Hypothesized Deep Features: Struct./Fun.	Kinetics	Thermo.	Equilib.
Reactions	J	F	A	L
Acid/Base	E	B	I	O
Molecules	N	D	K	H
Energy	G	P	M	C

Figure 1. (A) Sample question card "L". (B) Hypothesized Sorts. The columns represent Deep Feature groups and the rows represent Surface Feature groups. The highlighting shows that card "L" has "Reactions" as a surface feature and "Equilibrium" as a deep feature.

spanned traditional subdisciplines of chemistry, were presented with multiple representations, and were written to minimize jargon. Though there were a host of surface features we could have used, ours were chosen based on our experience teaching chemistry to first-year college students. For each question card, the embedded surface feature was explicit but superficial to the presented question. The biology card-sorting activity this work was modeled after aligned its hypothesized deep features to core concepts of biological literacy laid out in *Vision and Change*[20] and the curriculum framework for AP Biology.[21] The chemistry community has yet to find consensus around a parallel set of core chemistry literacy concepts like those in *Vision and Change*. We drew our hypothesized deep features from the AP Chemistry framework[22] and the *ACS Guidelines and Recommendations for the Teaching of High School Chemistry*.[23] This method is likely flexible enough to be used with different theoretical bases, and as a result, there were likely many deep feature sets we could have used. Ultimately, we chose these four categories because they crossed traditional chemistry subdisciplines and we believed that faculty in chemistry were likely to recognize them and likely to structure their understanding of chemistry around them.

To maintain the integrity of this instrument, the cards used in this study are not included for publication. Card sets are available upon request.

Sample

In Fall 2014, 418 students and 40 faculty from a large, comprehensive public university in the western United States were invited to participate in this study. The student pool was drawn from first-year, nonmajor general chemistry courses and targeted courses within the undergraduate Chemistry/Biochemistry curriculum. The faculty pool was the full-time

DOI: 10.1021/acs.jchemed.5b00992
J. Chem. Educ. XXXX, XXX, XXX−XXX

teaching faculty in a Department of Chemistry and Biochemistry. We focused on two populations, "novice" and "faculty," for validation purposes. On the basis of having earned Ph.D.'s in chemistry or a related field, we categorized faculty as putative experts. As noted before, expertise is not necessarily developed linearly with academic training[9] and we did not use other instruments to validate expertise, but we predicted that members of the faculty group would interact with the stimulus set by organizing around deep features, a proxy for expertise. This is consistent with identifying experts as having domain-specific knowledge in other categorization tasks.[12] A further limitation of this approach is that the faculty selected may not be representative of expertise in chemistry outside of this type of academic institution. To apply the tool to a cross-sectional analysis of a chemistry program, we included groups of first-year, second-year, and upper-division chemistry and biochemistry majors. See Table 1 for details on the sampled population.

Table 1. Distribution of Card-Sorting Participants from Each Population

Population	Description of Population	Participants, *N*
Novice	Nonchemistry or biochemistry majors in their first 2 weeks at the university	162
First Year	Chemistry or biochemistry majors in their first 2 weeks in our program	77
Second Year	Chemistry or biochemistry majors in the first quarter of their second year in our program	28
Upper Division	Chemistry or biochemistry majors with three or more years within our program	51
Faculty	Chemistry and biochemistry faculty with a Ph.D. in chemistry or a related field	31

Students who did not consent to participate (n = 2), did not complete the sorting activity following the given instructions (n = 9), or did not fit into one of four student categories described above (n = 89) were excluded from analysis. In total, 407 of 418 invited students completed the sorting activity, a 97% participation rate. Additionally, 31 of 40 invited faculty completed the sorting activity, a 78% participation rate.

Experimental Protocol

The card sorting activity was carried out in either a studio classroom setting[24] or in a computer lab during regular course meetings. Participants were provided with an informed consent document approved by the University's Institutional Review Board. Each individual sorted paper cards and entered his or her results on a secure Web site developed by the authors.

The first activity was an "unframed sort" (sometimes referred to as an "open sort" in the card sorting literature). Participants were prompted to consider what they knew about chemistry and to sort the chemistry question cards into no more than 15 groups that represent common underlying chemistry principles. They were asked to give each of their groups a name that described what the group represented to them. They were further instructed that there was no right or wrong way to group the cards and were allowed to work at their own pace.

The second activity was a "framed sort" (sometimes referred to as a "closed sort"). Participants repeated the sorting activity, this time sorting into four prenamed categories: Thermodynamics, Equilibrium, Structure/Function, Kinetics. These corresponded to the hypothesized deep feature categories embedded in the card set. Following each sorting activity, participants were prompted to answer open-ended questions about their sorting and to complete a short demographic questionnaire.

Edit Distance

The edit distance[18,25] (ED) was calculated as the minimum number of moves that would be required to make a participant's sort match one of the hypothesized sorts. For example, if a hypothesized sort was {ABC,DEF} and a participant's sort was {AB,CEF,D} it would require two moves to make the sorts match. With 16 cards and a hypothesized sort of four categories, the maximum edit distance in our implementation for any given participant's sort was 12. A low edit distance implies an individual sort is similar to a hypothesized sort. Thus, we can associate a lower ED to the hypothesized deep sort (ED-Deep) with more expert-like thinking. See the Supporting Information for details on how the edit distance was computed.

Percent Pairings

The number of card parings that were common between the hypothesized sorts and each participant's sort were counted. For example, if a participant's sort contained the grouping {ABC}, there were three possible pairings: "AB", "AC", and "BC." When counting the total number of pairs made by an individual, a card placed in a category by itself was counted as paired with a null card and was considered an unexpected pairing. The pairings were compared to those present in the hypothesized sorts to determine the number of deep feature

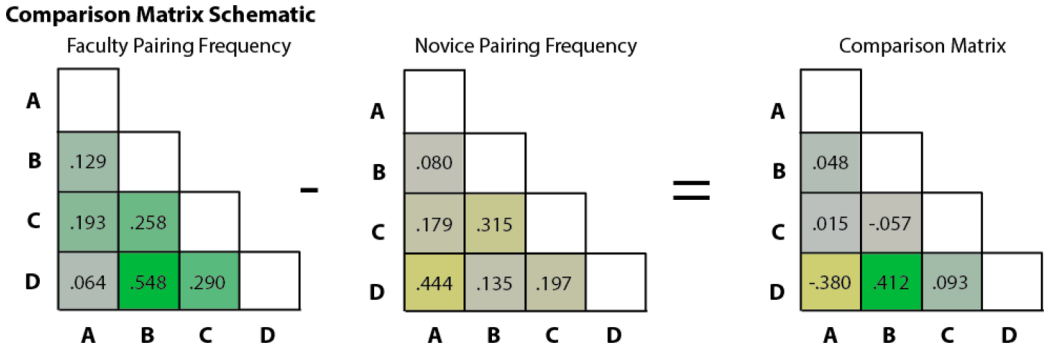


Figure 2. Comparison matrix construction from normalized pairing frequency matrices. The two left-most triangular matrices show the fraction of each population that paired each card together in their unframed sorts. The difference of these two matrices produces the comparison matrix, where positive values (shown in green) represent card pairings made more frequently by faculty than by novices and negative values (shown in yellow) represent card pairings made more frequently by novices.

DOI: 10.1021/acs.jchemed.5b00992
J. Chem. Educ. XXXX, XXX, XXX−XXX

pairings, surface feature pairings, and unexpected pairings. The total number of pairs varies considerably from sort to sort depending on the size of the groups present, so common pairings were normalized as percentages by dividing by the total number of pairs present in the sort. The more similar a given sort is to a hypothesized sort, the higher Percent Pairings (%P) value it will have. Thus, we can associate a larger %P in hypothesized deep sort with more expert-like thinking.

Hierarchical Cluster Analysis

Hierarchical cluster analysis was used to generate dendrogram plots illustrating the relationship between cards commonly sorted together by a population of participants. This technique generates a visualization of the clusters of cards sorted together experimentally by a population of participants without the need to compare to the hypothesized sorts generated by the researchers. To perform the cluster analysis, a normalized triangular pairing frequency matrix was generated for the novice and faculty populations (see Figure 2 and the Supporting Information). For each population, the matrix showed the frequency with which each card was paired with every other card. Hierarchical cluster analysis was performed on the pairing frequency matrix for the novice and faculty populations in the statistical software package JMP Pro.[26] Ward's minimum variance method[27] was used for the distance in the figures presented here, though the average linkage and centroid methods produced qualitatively similar results.

Comparison-Based Index

We developed a new metric to assign a single value to an individual participant's card sort that is independent of our hypotheses about how populations might sort. All of the other previously described numerical analytics compare to a single set of hypothesized deep and surface feature sorts, but the comparison-based index (CBI) compares against the distributions of actual expert and novice sorts without the need for hypothesized references.

A comparison matrix was built by subtracting the pairing frequency matrix of the novice population from the faculty population's pairing matrix. When building the comparison matrix, unpaired cards (i.e., cards that were put into categories by themselves) were also included. Figure 2 shows the construction of a portion of the comparison matrix. The leftmost matrix shows the fraction of the faculty population who paired each card together in the unframed sort, the middle matrix shows the fraction of the novice population who paired each card, and the comparison matrix is the difference between these matrices. Positive values in this matrix represent card pairings that were made more commonly by members of the faculty population than the novice population, whereas negative values represent pairings made more commonly by members of the novice population than the faculty population. For example, 54.8% of faculty paired cards B and D together, whereas only 13.5% of novices made the same pairing; the difference of these is given in the comparison matrix as the fraction 0.412 (the last digit varies because we calculated these values to higher precision than illustrated in the figure). Values close to zero in the comparison matrix either mean that neither population paired those cards together with high frequency or that they did so at a similar frequency, meaning that the pairing is not a good differentiator between the two populations.

For every card pairing that appeared in an individual's sort, the corresponding values from the comparison matrix were summed together to generate the CBI for the individual. For example, if an individual's sort contained the group {BCD}, the CBI for this portion of the sort would be the sum of the values looked up from the comparison matrix for the three card pairings that exist in this group: CBI = BC + BD + CD = −0.057 + 0.412 + 0.093 = 0.448. The more positive the value of the CBI, the more "expert-like" the sort was when referenced to this population of faculty and novices. It is important to note that the CBI is inherently empirical, developed with predefined populations of novices and expert and its utility comes in applying it to a new population.

Statistical Analysis

On multiple measures, *t*-tests were performed assuming unequal variances to determine whether novice and faculty populations were different. Although some distributions showed moderate skewness, the *t*-test is robust to non-normality for moderate to large sample sizes (Table 1). Furthermore, the degree of significance between the groups in each of these comparisons was large, as suggested by extremely small p values ($p < 0.0001$ for all comparisons), so it is unlikely that results would be changed by skewness in the data. An ANOVA with a Tukey−Kramer test was used to examine the variances between different populations for the CBI. Differences in the variances were small and the data within each group was either approximately normal or only slightly skewed. The sample sizes were moderate to large (Table 1) and differences were highly significant, so the Tukey−Kramer test is appropriate for comparing the populations.

Qualitative Analysis

Systematic coding analysis was applied to the category names assigned by participants in the unframed sort. Through an iterative process of coding to consensus by multiple researchers, a rubric with 19 codes was developed (see the Supporting Information). A rater coded the category names for groups of cards created by participants by assigning each category name to one or more of the codes in the rubric. Inter-rater reliability was estimated by double coding a pseudorandom sample of 5% of the category names (280) representing responses from each of the populations for every card. Two raters assigned identical codes to 86.1% of the category names. Cohen's κ[28] is a statistic often used to correct for the probability that agreement between two raters occurred by chance but cannot be applied here because category names were not restricted to a single code. Instead, we applied Mezzich's extension[29] of Cohen's κ, which allows for nonmutually exclusive coding. Using the proportional overlap procedure, the overall proportion of agreement was 0.878 and the proportion of chance agreement was 0.077, resulting in $\kappa = 0.868$. A κ value between 0.80 and 0.90 suggests a strong agreement between raters.[30]

RESULTS AND DISCUSSION

We focused our initial analysis on comparisons between novice ($n = 162$) and faculty ($n = 31$) sorters. The first comparisons use the edit distance and percent pairings described in the Methods section. Each of these measures relies on hypothesized deep features and hypothesized surface features. If the instrument generated valid data, we would expect that on multiple measures faculty sorters would be more likely to sort using deep features and less likely to sort using surface features, whereas novices would be more likely to sort using surface features and less likely to sort using deep features. We further employed analyses that did not reference our hypothesized features including hierarchical cluster analysis and a compar-

DOI: 10.1021/acs.jchemed.5b00992
J. Chem. Educ. XXXX, XXX, XXX−XXX

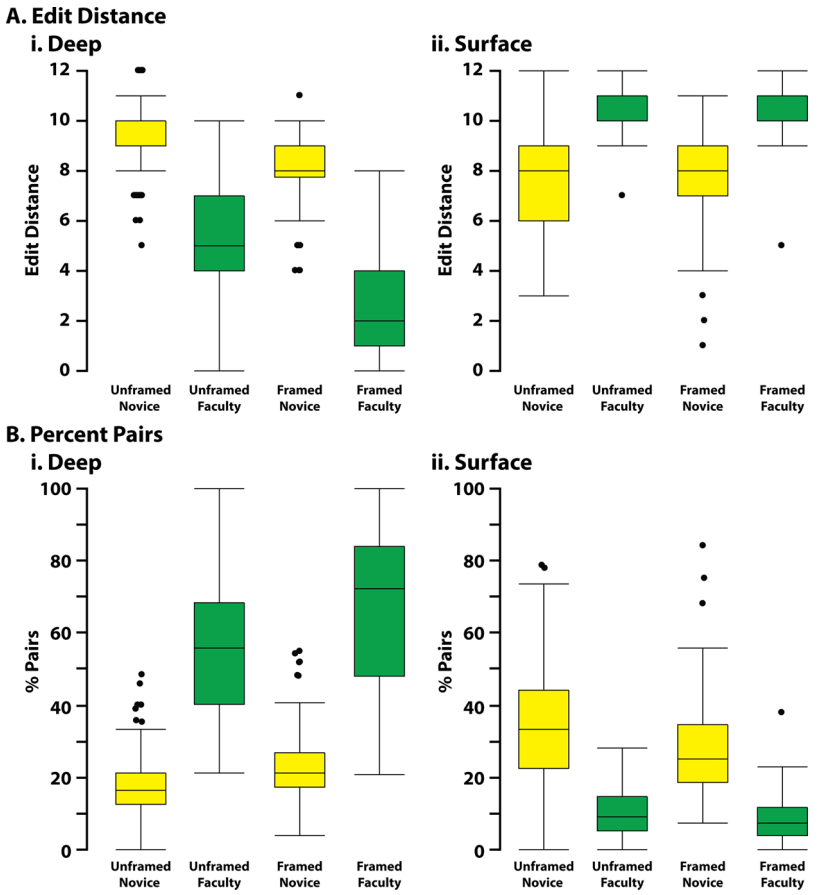


Figure 3. Comparisons of novice (yellow, n = 162) and faculty (green, n = 31) sorts (unframed and framed) relative to the hypothesized deep and surface sorts. Each box plot represents a statistically significant difference between novice and faculty sorts using a one tailed, t-test assuming unequal variance at 95% confidence, $p < 0.0001$ for each comparison. Effect sizes for these comparisons, estimated using Cohen's d, were large, ranging from 1.8 to 3.1. (A) Edit distances to the (i) hypothesized deep sort, and (ii) hypothesized surface sort. (B) Hypothesized (i) deep feature percent pairing and (ii) surface feature percent pairing.

Table 2. Average Values for Each Population on Edit Distance and Percent Pairs Metrics

		Deep Unframed		Surface Unframed		Deep Framed		Surface Framed	
Group	N	Mean	Standard Error	Mean	Standard Error	Mean	Standard Error	Mean	Standard Error
Novice: Average ED	162	9.41	0.10	7.44	0.15	8.18	0.11	7.49	0.13
Faculty Average ED	31	5.29	0.42	10.39	0.21	2.45	0.40	10.42	0.25
Novice Average % Pairs	162	17.4%	0.7%	34.8%	1.2%	22.8%	0.7%	27.5%	1.0%
Faculty Average % Pairs	31	55.7%	3.7%	10.1%	1.2%	69.6%	4.1%	8.7%	1.5%

ison-based index to directly compare the sorting behavior of different populations to each other.

Edit Distance

The edit distance (ED) reflects how many card moves one would have to make to move from a given sort to one of our hypothesized sorts (deep or surface). A low edit distance implies an individual sort is similar to a hypothesized sort. In the unframed sort, the average edit distance to the hypothesized deep feature sort (ED-Deep) for novice sorters was significantly greater than for faculty sorters, Figure 3Ai ($p < 0.0001$). Faculty sorters were more likely to have a smaller ED-Deep and, as a result, were more likely to sort using deep features than novices. The opposite was true for the average ED to the hypothesized surface feature sort (ED-Surface) as shown in Figure 3Aii. Novice sorters were significantly more likely to have a smaller ED-Surface, sorting with surface features more often than faculty ($p < 0.0001$). These results are similar in the framed sort (Figure 3Ai,ii). As with the unframed sort, faculty had a significantly lower average ED-Deep ($p < 0.0001$) and novices had a significantly lower average ED-Surface ($p < 0.0001$). Box plots in Figure 3a highlight the distribution of ED-deep values for faculty and novice populations. The ED-Deep values for the novice population overlaps with only the bottom quartile of faculty population, with the exception of a few outliers. For both sorting activities, faculty were more likely to have ED consistent with use of the embedded deep features, where novices were more likely to have ED consistent with use of the embedded surface features. Table 2 summarizes these findings.

Percent Pairing

Percent pairings (%P) (Figure 3B) offer a second metric for investigating how similar a given sort is to one of our hypothesized sorts (deep or surface). If a given sort is similar to a hypothesized sort, it will result in a relatively high %P value, representing that many of the pairs made in the sort were also

DOI: 10.1021/acs.jchemed.5b00992
J. Chem. Educ. XXXX, XXX, XXX−XXX

present in the hypothesized sort. In the unframed sort (Figure 3Bi), the average %P with the hypothesized deep feature sort (%P-Deep) for novice sorters was significantly lower than for faculty sorters ($p < 0.0001$). Faculty sorters were more likely to have a larger %P-Deep, hence more likely to be sorting using deep features than novices. Figure 3Bii shows that the average %P to the hypothesized surface feature sort (%P-Surface) for novice sorters was significantly higher than the average for the faculty population ($p < 0.0001$). While neither population had particularly high %P-Surface, novice sorters were significantly more likely to have a larger %P-Surface and thus more likely to be sorting using surface features than faculty. These significant differences are also apparent in the framed sort (Figure 3Bi,ii). In the framed sort, %P-Deep for novice sorters was again significantly lower than %P-Deep for faculty sorters, and the %P-Surface for novice sorters was significantly higher than %P-Surface for faculty sorters ($p < 0.0001$ for both). For both the framed and unframed sort, faculty were more likely to have %P metrics consistent with use of the embedded deep features while novices were more likely to have %P metrics consistent with use of the embedded surface features. Table 2 summarizes these findings.

Cluster Analysis

Figure 4 shows the dendrogram from the hierarchical clustering analysis from the unframed sort for the novice and faculty populations. The cluster analysis identifies clusters from the pairing frequency matrix that are similar and provides a visual tool for identifying cards frequently grouped by participants. The number of clusters was chosen at a natural break in a plot of the linkage distances joining each cluster versus the number of clusters (a scree plot) for the faculty population. The number of clusters in the novice population was chosen using the same linkage distance cutoff.

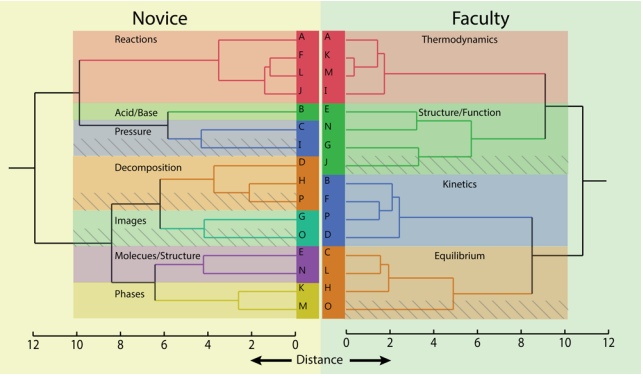


Figure 4. Hierarchical clustering analysis dendrogram for novice and faculty populations for the unframed sort. The distance is calculated using Ward's minimum variance method.[27] Shorter linkage distances indicate a higher affinity for grouping cards together. The most frequently used qualitative analysis code for each cluster is overlaid in the colored region; striped regions indicate that the label was the second most frequently used code (e.g., the most common code assigned to card O by faculty was "acid/base", but "equilibrium" was the second most frequently used code).

Despite the fact that the cluster analysis is blind to the hypothesized sorts, the four clusters in the faculty population match exactly with the hypothesized deep feature sort. This suggests that there is a good correlation between the hypothesized deep feature sort and the way that faculty tend to sort. The faculty did not all group their cards to match this clustering scheme (i.e., there were a variety of ways that the faculty sorted), so it is useful to look at how closely linked each of the cards in the clusters are. The cards in the "Thermodynamics" cluster were grouped together frequently by the faculty population as indicated by the short linkage distances in the dendrogram. The most frequently used category names by faculty (overlaid in Figure 4) for each of these clusters suggests that the majority of faculty were in fact sorting based on the embedded deep features rather than using an alternative theoretical framework.

The novice dendrogram is less definitive. Many more clusters appeared than for the faculty at the same level of similarity and tended to have greater distance between clustered cards. This suggests that novices were not uniformly identifying the same sets of surface features hypothesized during the card set construction. Though the faculty and novice populations both produced a distribution of card sorts, this shows that the faculty were more likely to converge on similar features. Inspection of the novice clusters in parallel with the qualitative analysis suggested that there were other unanticipated surface features present in the card set. For example, though not intentionally embedded as a surface feature, the word "decomposition" appeared in several of the problems and many novices grouped these cards together.

In the unframed sort, participants were asked to give each card group they created a name that described the underlying chemistry principle that led them to form that particular group. Systematic coding analysis of group names allowed us to identify the top faculty group codes and top novice group codes for each card. The most frequently used codes by each population are the cluster names overlaid in Figure 4. Table 3 shows how many of the participants from each population applied the most frequently used faculty and novice codes. For 14 of the 16 cards, the top codes for faculty and novices are different. As an example, for card L, 71% of faculty used the top faculty code, "Equilibrium", whereas no novices used this code. In contrast, for card L, 68% of novices used the top novice code, "Reactions", whereas only 3% of faculty used this code. For most cards, there was large difference between faculty and novices. In two cases, card J and card O, the top codes for faculty and novices were the same. Not coincidently, in the dendrogram in Figure 4, these are the only two cards where the faculty's second most used card names matched the cluster

Table 3. Comparison of Top Code Usage for Faculty and Novice Sorters for Each Question Card

		Top Code Usage by Card for Faculty and Novices, %															
Approach	Population	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Use of Top Faculty Code	Faculty	71	45	61	74	55	48	32	61	55	39	65	71	65	55	48	52
	Novices	0	4	0	1	6	0	1	0	1	65	2	0	2	4	40	1
Use of Top Novice Code	Faculty	3	10	0	0	16	39	0	0	0	39	23	3	26	19	48	0
	Novices	44	45	22	40	24	68	33	36	25	65	34	68	33	23	40	33

DOI: 10.1021/acs.jchemed.5b00992
J. Chem. Educ. XXXX, XXX, XXX−XXX

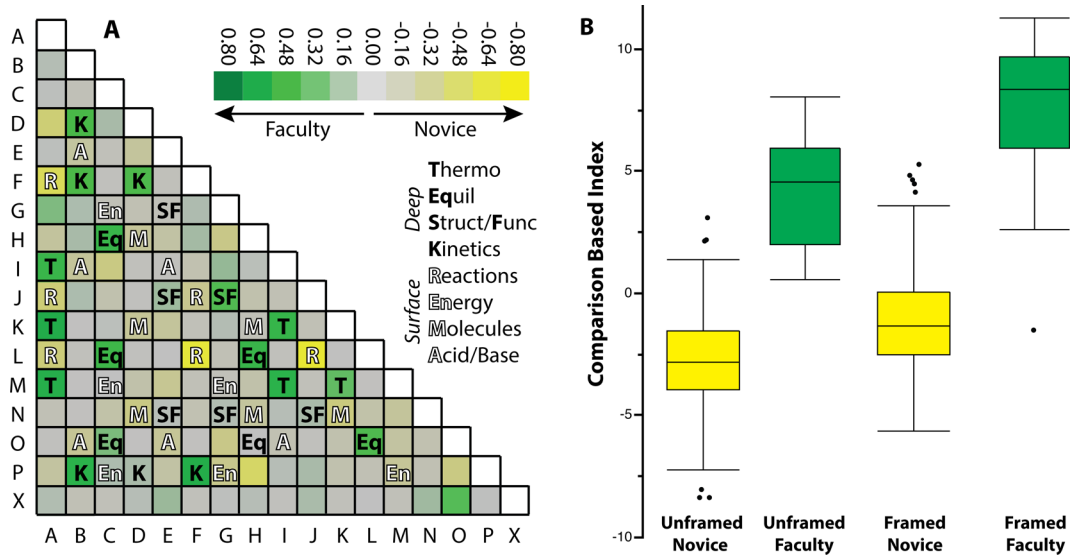


Figure 5. (A) Comparison matrix for the unframed sort and (B) Comparison based index (CBI) for novice and faculty populations. The comparison matrix shows those card pairings which best differentiate the faculty and novice populations. The X represents cards that were left unpaired. Hypothesized deep and surface feature pairings are overlaid. The unframed CBI shown in (B) were calculated from this comparison matrix.

labels. For some cards, like G for both faculty and novices, C for novices, and J for faculty, the top codes were used less frequently. It may be that the embedded hypothesized deep or surface features were not as easily identifiable. Looking at Card G as an example, only 32% of faculty used the top faculty code and only 33% of novices used the top novice code. In essence, Table 3 gives an idea of the effectiveness of each card. Cards where faculty and novices assigned different codes with a high frequency are more effective differentiators.

Comparison-Based Index

Although the edit distance and percent pairings are effective at measuring differences between faculty and novice populations of sorters, they rely on the experimenter's preconceived hypotheses about how those populations might sort. As demonstrated in the cluster analysis above, there is no single novice or faculty sort, but there are some common approaches taken to sorting by members of those populations. To provide an empirical measure of how expert- or novice-like an individual sort is, we introduce a new tool, the comparison-based index (CBI), which provides a single number for a sort compared against the actual sorting characteristics of a population of faculty and novices.

Figure 5 shows the CBI for the novice and faculty populations. The unframed CBI values shown in Figure 5B were calculated from the comparison matrix represented in Figure 5A (see the Supporting Information for the framed comparison matrix). The comparison matrix provides a visual representation of the pairings that were made more frequently by faculty (green) and by novices (yellow) and is overlaid with the hypothesized deep and surface features. Cells with distinct color without a hypothesized category overlaid represent unexpected pairings that occurred with high frequency.

Several features are apparent in this color-coded comparison matrix that suggest the value of using a metric that is independent of hypothesized sorts. There are fewer dark yellow than dark green cells in the matrix, indicating that the novice sorts were more evenly distributed across the available sorting space, whereas the faculty tended to converge on a consistent group of pairings in their sorts. This is consistent with the hierarchical clustering data of Figure 4. The cells that were distinctly "faculty" correlate strongly with the hypothesized deep features; the most prominent exception was card O, which was left unpaired by 38.7% of faculty and only 4.3% of novices. In contrast, the distinctly "novice" pairings did not correlate as strongly with hypothesized features because the novices identified other surface features not intentionally built into the card set. Although the percent of unexpected pairs leads to this same conclusion, these matrices give insight into what those unexpected pairs were.

An example of the insight that can be gained from inspection of the qualitative analysis for unexpected pairs can be found in the group {ADHP}. All of these cards contained the word "decomposition." All four of those cards were placed in categories coded as "decomposition" with a frequency between 30 and 40% by the novice population while the members of the faculty population never used the code at all. Although not included as a hypothesized surface feature by design in the card set, decomposition was a true surface feature that arose from the sorting task. Card A does not appear to be grouped with D, H, and P in Figure 4 because A was more frequently associated with the hypothesized surface feature "reactions", but "decomposition" was the second most frequently used code for this card. This would not be as readily evident from the cluster analysis, which demonstrates the utility of the pairing frequency matrices. Additionally, the card pairing HP was the most frequently paired within this grouping. Both of these cards were coded as "decomposition" or "images" with similar frequency by novices. Multiple approaches to sorting by novices captured these cards as a pairing based on features that were entirely ignored by faculty.

Although both the cluster analysis and inspection of the pairing matrices reveal that faculty tend to organize around the hypothesized deep features, there was still a distribution of ways of sorting the cards by the faculty population. The CBI offers a way to measure an individual's sort against the distribution of ways that faculty and novices in the field will organize the problems presented in the card set. Figure 5B shows the CBI for the novice and faculty populations. Because the CBI is built

DOI: 10.1021/acs.jchemed.5b00992
J. Chem. Educ. XXXX, XXX, XXX−XXX

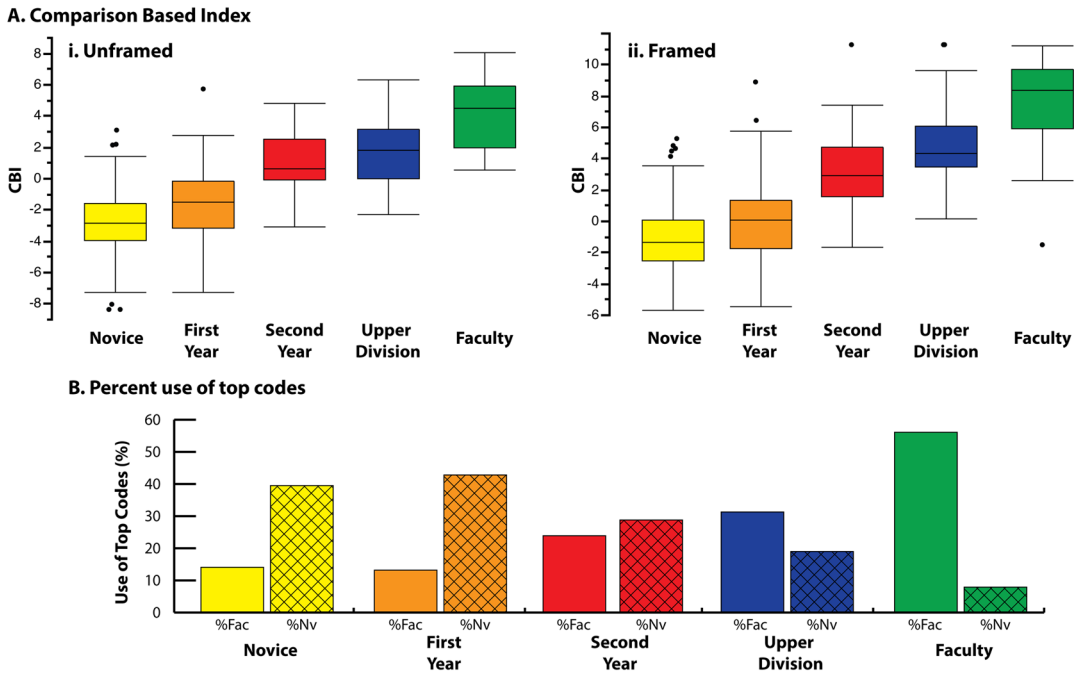


Figure 6. Departmental snapshot based on the unframed and framed sort. Comparisons of novices (yellow, n = 162), first year majors (orange, n = 77), second year majors (red, n = 28), upper division majors (blue, n = 51), and faculty (green, n = 31). (A) Comparison-based index (CBI) values for (i) unframed and (ii) framed sorts. (B) Comparison of top codes of each population in departmental snapshot. %Fac and %Nv refer to the percent of top faculty and novice codes applied to category names for each card (see Table 3). For the unframed sort (Ai), in an ANOVA with a Tukey−Kramer test, each of the populations are significantly different from each other on CBI (p < 0.0001) except for the second year-upper division comparison. Similarly, for the framed sort (Aii) each of the populations are significantly different from each other for all comparisons on CBI (p < 0.020). For both the unframed sort and framed sort the effect size was large (η^2 = 0.57 and 0.60 for unframed and framed, respectively).

from a matrix using the faculty and novice sorts as a training set, the faculty will, by definition, have a higher CBI than the novices, but this measure becomes particularly useful when comparing other populations to these faculty and novice populations (see the Application section below). Moreover, when the deep-feature categories were introduced in the framed sort, the CBI increased for both the novice and faculty populations. The average faculty CBI increased by almost twice as much as the novice CBI. The relative magnitude of the increase in CBI indicates that the frame led to a more narrow distribution of sorts for the faculty than it did for the novices, which suggests that faculty recognized the hypothesized deep features in the cards based on the framed categories better than novices did.

Application to Chemistry Program Assessment

In an effort to create a departmental "snapshot" relevant for program assessment, we focused on comparisons between novice (n = 162), first year majors (n = 77), second year majors (n = 28), upper division majors (n = 51), and faculty (n = 31). The entire task took an average 26.1 ± 0.5 min, with no significant difference between any population; 80% completed between 16.1 and 37.1 min. Figure 6A reports unframed comparison-based indexes (CBI) for each population. For both metrics, there was a trend toward more expert-like sorting as students progress through our program. Somewhat surprisingly, we were able to detect significant differences between novice and first year sorters (p < 0.0001). Because the novice and first-year groups are both in their first few weeks of study, this difference is almost certainly not a result of the chemistry program at the university. It likely reflects a difference in the population who chose a chemistry or biochemistry major. As illustrated in Figure 6A, the novice and first-year populations are similar except for the upper extreme on the CBI for the first-year majors.

The biggest jump toward expert-like sorting occurred between first and second year majors (ΔCBI = 2.7). Although we do not have the analytical power in this data set to statistically differentiate between second year and upper division students, the trend is an increase in expert-like sorting as students advance through our program. Framed CBI values for each population (Figure 6Aii) also showed significant differences and followed the same general trend. Figure 6B highlights how different populations in our departmental snapshot use the top novice and top faculty codes when naming categories. First year students use more novice-like codes (fewer faculty-like codes) than second year students, and second year students use more novice-like codes (fewer faculty-like codes) than upper-division students. Taken together with the CBI trend, this evidence suggests students are developing more expert-like thinking as they progress within our program.

The growth toward more expert-like conceptual thinking observed in this data set suggests that this card sorting task could provide useful information in program assessment. With larger populations of intermediate-level students, longitudinal tracking of individual students, and analysis of follow-up questions to the sorting task, we should be able to identify in what areas related to the task conceptual growth is occurring, where it is not, and how that is linked to coursework and related experiences.

Framed Vs Unframed Sorts

Subjects were asked to perform an unframed sort where they sorted using their own categories, and a framed sort where

DOI: 10.1021/acs.jchemed.5b00992
J. Chem. Educ. XXXX, XXX, XXX−XXX

subjects sorted cards into four prenamed categories that corresponded to the hypothesized deep feature categories. On the basis of average edit distances and average %Pairs to the hypothesized deep sort (Figures 2Ai and 2Bi), framing increased the probability that subjects sorted based on the embedded hypothesized deep features. Although this was true for both faculty and novices, the effect was more pronounced with faculty sorters. As an example, looking at changes in % Pairing between framed and unframed sorts (Figure 2bi), the novice population increases from an average of 17.4% ± 0.7% to 22.8% ± 0.7% deep pairs, a net gain of 5.4%, whereas the faculty population increases from 55.6% ± 3.7% to 69.6% ± 4.1% deep pairs, a net gain of 14.0%. Perhaps faculty sorters, who may have used a different strategy for their unframed sort, were more likely to recognize the hypothesized deep feature categories as a viable alternative. The result was that the framed sort offers a greater contrast between faculty and novices. These trends were also seen in Figure 6 for our departmental cross section. Looking at average CBI values for the unframed and framed sorts (Figure 6A), we could detect a statistically significant difference between second year students and upper division students with the framed sort, which did not appear in the unframed sort.

CONCLUSIONS

In this study, participants were asked to complete unframed and framed sorts of a set of 16 chemistry question cards, embedded with hypothesized deep and surface features, to answer our first research question: can a simple card sorting exercise distinguish between populations of novices and putative experts in chemistry? Based on multiple metrics tied to hypothesized deep and surface features, both of these card sorting exercises were able to statistically differentiate between populations of faculty and novice sorters (Figure 3). This relatively simple card-sorting task is an instrument that generates valid data for distinguishing between novices and putative experts.

By carefully examining the actual card-pairs made by faculty and novice sorters, we were able to address our second research question: is there an effective way to measure how expert-like an individual card-sorting result is without measuring against an experimenter's expectation of how experts will organize information? Edit distance and percent pairing both rely on the hypothesized deep and surface features embedded in the question cards, which in turn rely to some degree on experimenters' expectations. As is clear from the dendrogram in Figure 4 and to a lesser extent the Deep/Surface comparisons in Figure 3, as experimenters, we were better at predicting expert-like behavior than novice-like behavior. Faculty were drawn more strongly to deep features than novices were drawn to surface features, consistent with literature definitions of expertise described in the introduction. The new comparison-based index (CBI) metric relied on actual card-pairs made by faculty and novice populations, not experimenters' preconceptions. This new metric is as effective at differentiating (Figure 6) as our metrics that rely on hypothesized sorts, but is hypothesis-independent. With an experimental novice−faculty pair comparison matrix in hand, this card sorting activity has yet another metric to characterize student growth.

With hypothesis- and comparison-driven metrics in place and sorting data for a cross section of our departmental student population, we gained some insight into our third research question: what can the sorting task tell us about the development of expert-like thinking through a chemistry education curriculum as a tool for program assessment? Figure 6 provides a proof of principle that the task measures significant growth on comparison driven metrics. Although we established a set of methods here, the data suggest that insights regarding the use of various theoretical frameworks and their relationship to the curriculum may be obtained in future studies, though these analyses are beyond the scope of this paper. To be an effective longitudinal tool, this would have to be a repeatable measure. Similar to studies in biology, the framed sort may be more effective at distinguishing between putative expert and novice sorters.[18] Though the unframed sort may have a smaller contrast, it may be a more appropriate tool for longitudinal studies as it can be administered multiple times without exposing the hypothesized deep feature categories.

In short, we have introduced an instrument (card sets are available upon request) that can probe conceptual expertise in chemistry through a simple card sorting activity. We have also developed novel metrics that are hypothesis-independent. In the future, we plan to use this tool in a longitudinal study to track the development of expert-like thinking in students within our program.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available on the ACS Publications website at DOI: 10.1021/acs.jchemed.5b00992.

The Supporting Information includes additional experimental details, pairing frequency matrices, details on the qualitative analysis, and analysis code. (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: gscott02@calpoly.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors are grateful for participation by students and faculty. S.D.B. acknowledges his institution for sabbatical leave for this project. A National Science Foundation CAREER Award #0954127 supported K.D.T. and her research colleague Dr. Julia Smith of Holy Names University in the development of the Biology Card Sorting Task, in which this research has its origins. We thank John Walker in the Statistics Department at California Polytechnic State University for helpful consultations on statistical tests. We are also indebted to members of SEPAL: Science Education Partnership and Assessment Laboratory in the Department of Biology at San Francisco State University for helpful comments and discussion, in particular Dr. Sarah Bissonnette and Mr. John Rodriguez.

REFERENCES

(1) Lythcott, J. Problem Solving and Requisite Knowledge of Chemistry. *J. Chem. Educ.* **1990**, *67* (3), 248.

(2) Claesgens, J.; Daubenmire, P. L.; Scalise, K. M.; Balicki, S.; Gochyyev, P.; Stacy, A. M. What Does a Student Know Who Earns a Top Score on the Advanced Placement Chemistry Exam? *J. Chem. Educ.* **2014**, *91* (4), 472−479.

DOI: 10.1021/acs.jchemed.5b00992
J. Chem. Educ. XXXX, XXX, XXX−XXX

(3) Bodner, G. M. I Have Found You an Argument: The Conceptual Knowledge of Beginning Chemistry Graduate Students. *J. Chem. Educ.* **1991**, *68* (5), 385.
(4) Mulford, D. R.; Robinson, W. R. An Inventory for Alternate Conceptions among First-Semester General Chemistry Students. *J. Chem. Educ.* **2002**, *79* (6), 739.
(5) Claesgens, J.; Scalise, K.; Wilson, M.; Stacy, A. Mapping Student Understanding in Chemistry: The Perspectives of Chemists. *Sci. Educ.* **2009**, *93* (1), 56−85.
(6) Pendley, B. D.; Bretz, R. L.; Novak, J. D. Concept Maps as a Tool To Assess Learning in Chemistry. *J. Chem. Educ.* **1994**, *71* (1), 9.
(7) Neiles, K. Y. Measuring Knowledge: Tools To Measure Students' Mental Organization of Chemistry Information. In *Tools of Chemistry Education Research*; Bunce, D. M., Cole, R. S., Eds.; ACS Symposium Series; American Chemical Society: Washington, DC, 2014; Vol. *1166*, pp 169−189.
(8) Bédard, J.; Chi, M. T. H. Expertise. *Curr. Dir. Psychol. Sci.* **1992**, *1* (4), 135−139.
(9) Stains, M.; Talanquer, V. Classification of Chemical Reactions: Stages of Expertise. *J. Res. Sci. Teach.* **2008**, *45* (7), 771−793.
(10) Sevian, H.; Talanquer, V. Rethinking Chemistry: A Learning Progression on Chemical Thinking. *Chem. Educ. Res. Pract.* **2014**, *15* (1), 10−23.
(11) Chi, M. T. H.; Feltovich, P. J.; Glaser, R. Categorization and Representation of Physics Problems by Experts and Novices. *Cogn. Sci.* **1981**, *5* (2), 121−152.
(12) Rottman, B. M.; Gentner, D.; Goldwater, M. B. Causal Systems Categories: Differences in Novice and Expert Categorization of Causal Phenomena. *Cogn. Sci.* **2012**, *36* (5), 919−932.
(13) Hudson, W. Card Sorting. In *Encyclopedia of Human Computer Interaction*; Soegaard, M., Dam, R. F., Eds.; The Interaction Design Foundation: Aarhus, Denmark, 2012.
(14) Singh, C. Categorization of Problems to Assess and Improve Proficiency as Teachers and Learners. *Am. J. Phys.* **2009**, *77* (1), 73.
(15) Mason, A.; Singh, C. Assessing Expertise in Introductory Physics Using Categorization Task. *Phys. Rev. ST Phys. Educ. Res.* **2011**, *7* (2), 020110.
(16) Wolf, S. F.; Dougherty, D. P.; Kortemeyer, G. Rigging the Deck: Selecting Good Problems for Expert-Novice Card-Sorting Experiments. *Phys. Rev. ST Phys. Educ. Res.* **2012**, *8* (2), 020116.
(17) Lin, S.-Y.; Singh, C. Categorization of Quantum Mechanics Problems by Professors and Students. *Eur. J. Phys.* **2010**, *31* (1), 57−68.
(18) Smith, J. I.; Combs, E. D.; Nagami, P. H.; Alto, V. M.; Goh, H. G.; Gourdet, M. A. A.; Hough, C. M.; Nickell, A. E.; Peer, A. G.; Coley, J. D.; et al. Development of the Biology Card Sorting Task to Measure Conceptual Expertise in Biology. *CBE Life Sci. Educ.* **2013**, *12* (4), 628−644.
(19) Kozma, R. B.; Russell, J. Multimedia and Understanding: Expert and Novice Responses to Different Representations of Chemical Phenomena. *J. Res. Sci. Teach.* **1997**, *34* (9), 949−968.
(20) Woodin, T.; Carter, V. C.; Fletcher, L. Vision and Change in Biology Undergraduate Education, a Call for Action–Initial Responses. *CBE Life Sci. Educ.* **2010**, *9* (2), 71−73.
(21) College Board. *AP Biology Curriculum Framework 2012−2013*; College Board: New York, NY, 2011.
(22) College Board. *AP Chemistry Course and Exam Description*; College Board: New York, NY, 2014.
(23) American Chemical Society. *ACS Guidelines and Recommendations for the Teaching of High School Chemistry*; The American Chemical Society: Washington, DC, 2012.
(24) Bailey, C.; Kingsbury, K.; Kulinowski, K.; Paradis, J.; Schoonover, R. An Integrated Lecture-Laboratory Environment for General Chemistry. *J. Chem. Educ.* **2000**, *77* (2), 195.
(25) Deibel, K.; Anderson, R.; Anderson, R. Using Edit Distance to Analyze Card Sorts. *Expert Syst.* **2005**, *22* (3), 129−138.
(26) *JMP*, Version 11.2; SAS Institute Inc.: Cary, NC, 1989−2007.
(27) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58* (301), 236−244.
(28) Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20* (1), 37−46.
(29) Mezzich, J. E.; Kraemer, H. C.; Worthington, D. R.; Coffman, G. A. Assessment of Agreement among Several Raters Formulating Multiple Diagnoses. *J. Psychiatr. Res.* **1981**, *16* (1), 29−39.
(30) McHugh, M. L. Interrater Reliability: The Kappa Statistic. *Biochem. medica* **2012**, *22* (3), 276−282.

DOI: 10.1021/acs.jchemed.5b00992
J. Chem. Educ. XXXX, XXX, XXX−XXX