Differentiating expert and novice cognitive structures

Steven F. Wolf[*,†], Daniel P. Dougherty[†] and Gerd Kortemeyer[†,*]

[*]*Department of Physics and Astronomy Michigan State University, East Lansing, MI 48824, USA*
[†]*Lyman Briggs College Michigan State University, East Lansing, MI 48825, USA*

Abstract. A seminal study by Chi *et al.* firmly established the paradigm that novices categorize physics problems by "surface features" (e.g. "incline," "pendulum," "projectile motion," ...), while experts use "deep structure" (e.g. "energy conservation," "Newton 2," ...). Yet, efforts to replicate the study frequently fail, since the ability to distinguish experts from novices is highly sensitive to the problem set being used. But what properties of problems are most important in problem sets that discriminate experts from novices in a measurable way? To answer this question, we studied the categorizations by known physics experts and novices using a large, diverse set of problems, in order to subsequently study how well these two groups can be discriminated using small subsets. Having a large initial set allowed us to form a large number of smaller subsets and study their properties. We found that the number of questions required to accurately classify experts and novices could be surprisingly small so long as the problem set is carefully crafted to be composed of problems with particular pedagogical and contextual features.

Keywords: Categorization, Graph Theory, Cognitive Structure, Expert and Novice
PACS: 01.40.Fk,01.40.Ha,01.55.+b,01.90.+g

INTRODUCTION

The physics education research community has recently been grappling with different understandings of student learning and the conceptualization of expertise. The predominant paradigm that the PER community has used to understand expertise was established by the seminal study done by Chi et al. [1]. However, replicating this experiment in a way which clearly distinguishes between experts and novices is challenging. More often than not, attempts to verify it fail, as an informal survey among physics education researchers indicates [2]. In order to understand why this occurs, we have designed and carried out a categorization experiment and developed a novel methodology to analyze that experiment [2].

Our methodology is unique in that it describes single categorizations as graphs, quantifies the difference between any two categorization graphs using our distance metric, and visualizes the relative position of each of the sorters using Principal Components Analysis (PCA) [2]. PCA is used to analyze multivariable data, effectively performing a rotation on that data so that the highest degree of variability is contained in a small number of dimensions. Our use of PCA is analogous to placing cities (in our case: sorters) on a map given the driving distances (in our case: distances calculated by our metric) between those cities.

Our first paper focused on what properties of categorizations would discriminate experts from novices [2]. We found that this discrimination is evident only when categorizations are compared on a microscopic–problem specific–basis rather than macroscopic features of the categorizations [2]. This finding suggests that the experiment of Chi et al. is difficult to replicate because the largest source of variation is not expertise. We now explore *why* a particular set of problems would discriminate experts from novices while others do not.

The largest source of variation determined by the PCA—first principal component—is related to the sorting behavior, something we termed "stackers" and "spreaders," rather than expertise [2]. In other words, the most dominant characteristic of the sorters' categorizations has nothing to do with being an expert or a novice. Only the second principal component differentiates experts from novices. Interpreting the PCA is a difficult task given our implementation of the algorithm. Previously, we were able to interpret the first principal component because of the results of our Cognitive Categorization Model [2]. In order to interpret the second principal component, we hypothesized that there are two sources of this variability, namely the pedagogical and contextual features of the problems and distinct internal cognitive processes of each sorter. For now, we will focus on the problem-specific nature of sorter discrimination rather than internal cognitive processes.

Variability in student performance while working on problems in relativity [3] and motion [4] is well-documented. It is reasonable to believe that this problem-dependent nature of student reasoning extends from solving problems to the categorization behavior of sorters. By studying the level of discrimination between experts and novices for many subsets of problems from our earlier experiment [2], we investigated the extent to which the different pedagogical and contextual features of the

problems can account for expert-novice discrimination.

METHODOLOGY

Previously, we designed and carried out a card-sorting experiment on physics experts and novices at Michigan State University, adhering to the experimental method of Chi et al. as closely as possible [2]. A total of 18 physics professors and 23 novices participated in our study. All of the novices had completed at least the first semester of an introductory physics course at MSU. We gave each sorter a set of 50 problems to sort based on "similarity of solution," explicitly following the prompt of Mason and Singh [5]. Each sorter categorized his or her problems and recorded the groups and group names in a separate packet. Multiple categorization—putting a single card into more than one category—was allowed, but not expected.

We have chosen such a large set of cards in order to study the properties of subsets. We constructed a set of problems that was diverse in terms of both content and cognitive demands. In order to do this, we considered two traditional measures, the chapter which a problem is in (using Walker's textbook [6] as a guide) and the problem difficulty as measured by the number of "dots" a problem has. We also included the taxonomic classification according to the Taxonomy of Introductory Physics Problems (TIPP) [7]. The TIPP is useful because it considers two dimensions or knowledge domains, one for declarative knowledge (information) and the other for procedural knowledge (mental procedures) [7]. See Table 1 for a list of the TIPP levels included in our study. Higher levels were not included because they are more suited to research projects rather than homework problems [7]. Each problem was therefore classified with four statistics: The highest complex cognitive process that is necessary to solve it for both declarative knowledge (TIPP-D) and procedural knowledge (TIPP-P), the problem difficulty (DIFF), and finally the chapter that the problem could be found (CHAP).

What is the composition of a minimal ideal subset? What characteristics need to be present? In other words, instead of picking random problems from the back of chapters in textbooks, how much does a problem set need to be "rigged" in order to be effective in discriminating experts and novices. As a proof of concept, we have compared the sorter visualizations from the entire dataset from our previous study [2] to the subset of problems within that set which we obtained from Singh's study [8]. As can be seen in Figure 1 (a) and (b), the two visualizations have similar properties and are both interpreted as we discussed in the previous section. However, the details of these figures are different. Sorters have different relative positions. This shows that this analysis and visualization method is indeed sensitive to the problems consideration. As both of these problem sets have similar properties, the overall level of expert-novice discrimination was also similar.

TABLE 1. A limited hierarchy of the cognitive processes described by the TIPP. Most problems in a standard physics textbook require a highest declarative knowledge process of Comprehension–Integrating, and procedural knowledge process of Retrieval–Executing [7]. The level indicates the numeric value we scored the highest cognitive process required by each problem.

Cognitive Process	Sub-process	Level
Retrieval	Recall/Recognize	1
	Executing*	2
Comprehension	Integrating	3
	Symbolizing	4
Analysis	Matching	5
	Classifying	6
	Analyzing Errors	7

* Procedural Knowledge only

In order to study the relationship between the problem properties and the degree of discrimination between experts and novices, we analyzed 40000 5-problem subsets and 275000 random 10-problem subsets of our original 50-problem set. As these were random subsets, they were quite diverse in terms of the properties of the problems within each subset and varied a great deal in the power to discriminate between the different kinds of sorters. Our method enables us to study a large number of problem sets with a limited number of human sorters, but we are aware that this is not fully equivalent. This experiment is limited to looking at patterns within the data already in our set.

Our initial experiment found that experts and novices were separated by the second principal component. However, this will not be sufficient to quantify the differentiation between expert and novice groups. Instead, we used three statistical tests to characterize the discrimination between these groups. These are the Hotelling's test [9], Cramer test [10], and Average Rate of Correct Classification (ARCC) [11]. The Hotelling's test is a standard test used to compare two groups of multivariable data, which assumes that each group's distribution of these data is elliptical [9]. The Cramer test is a non-parametric analog of this test, that is, it does not assume a distribution shape [10]. The ARCC is a statistical test which relies on Linear Discriminant Analysis to determine how well experts and novices are separated by the PCA.

In order to determine which of the problem properties are most important, we combined all of these measures into a Canonical Correlation Analysis (CCA) [12].

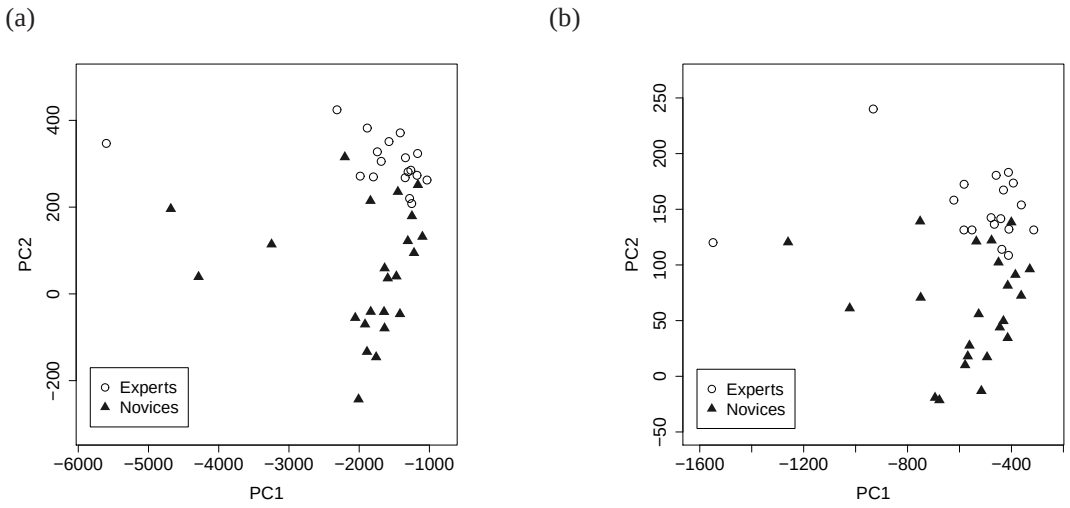

FIGURE 1. (a) This is the Principal Components Analysis (PCA) plot of the sorters for the entire set of problems from our previous study [2]. Both a Cramer's test ($p = 0.048$) and a Hotelling's test ($p < 10^{-5}$) find the expert and novice groups to be distinct at a 95% confidence level. (b) This is the PCA plot of the sorters considering only the problems from Singh's study. Both a Cramer's test ($p = 0.041$) and a Hotelling's test ($p < 10^{-5}$) find the expert and novice groups to be distinct at a 95% confidence level. For both plots, PC1 is the coordinate along the first principal axis, and PC2 is the coordinate along the second principal axis. Sorters known by us to be experts are marked by circles while sorters known by us to be novices are marked by filled triangles. The second principal component discriminates experts from novices.

A CCA uses the relationship between the predictor variables (in our case the problem statistics) to the explanatory variables (the sorter discrimination statistics). Finally, we found the variability explained by each group of problem statistics (e.g. all of the TIPP-P statistics) in this manner:

$$\mathrm{var}_{\mathrm{stat}} = \frac{\mathrm{cor}_{\mathrm{stat}}}{\sum_{\mathrm{stats}} \mathrm{cor}_{\mathrm{stat}}} \tag{1}$$

where $\mathrm{cor}_{\mathrm{stat}}$ is the correlation coefficient found by calculating a CCA for each group of problem statistics with the sorter discrimination statistics and the summation in the denominator is done over the groups of problem statistics.

RESULTS

In our analysis of the subsets of problems, we found that the ability of these subsets to distinguish experts from novices varied from negligible levels to nearly total separation. Moreover, this behavior was prevelant for both the 5-problem and 10-problem subsets. Because CCA assumes linear relationships between the predictor variables and the response variables, we did a log transformation on all of our variables in order to linearize power-law relationships and symmetrize skewed distributions. For the 5-problem subsets, we found a correlation coefficient of $r^2 = 0.359$, while for the 10-problem subsets, we found $r^2 = 0.427$. Given this result, it is clear that there should be a problem set size effect on the ability to discriminate experts from novices. However caution is warranted: The Cognitive Categorization Model developed in our previous study showed that sorters choose a number of categories based on the number of problems that are in the problem set [2]. It is highly unlikely that a sorter would choose to put all cards in a single category, nor is it likely that a sorter would create one category per problem. However, random subsets of problems chosen from a large set of problems could exhibit these extreme behaviors quite frequently.

Investigating the relationships found by the CCA further, we found that the most variability is contained in the Chapter variables, followed by the TIPP-P variables (See Table 2). The fact that Chapter variables explained a great deal of variability is not surprising since this is our analog for "deep structure" (i.e. there is a Force chapter, an Energy chapter, a Momentum chapter, etc.). However, more interesting is the prominence of TIPP-P in explaining expert-novice sorting differences. Therefore, it is important that the problem set under consideration asks questions that require more than calculation and include tasks such as making a flow chart of a problem solving strategy. It is possible that the prominence of the TIPP-P statistic is due in part to the nature of the students at MSU. As the physics courses are in a large lecture format and require computerized homework, questions that require hand-grading are few. Therefore these sorts of problems may have "surprised" our novice sorters, which

TABLE 2. Variability explained by each of our problem variable groups among our 10-problem subsets. From this we see that the Chapter is an important variable, followed by the TIPP-P statistic.

Problem variable group	Percent variability explained
Chapter	41.4
TIPP-P	30.4
Difficulty	22.8
TIPP-D	5.4

may have affected their ability to sort these problems. Problem difficulty was the next most important statistic. Here we found that the "easy" problems—as determined by a typical textbook author—were the most important in discriminating expert from novice. However, it is possible that these problems were deceptively easy, or that the novices were over-thinking these problems. Table 2 also clearly indicates that the TIPP-D level does not play a large role in discriminating experts from novices.

CONCLUSION

As was found by Chi et al., we agree that deep structure is an important feature determining the difference between experts and novices. For this reason it is important to construct a problem set with problems from a variety of chapters–our analog for "deep structure". Yet, the frequent null results obtained when replicating this experiment, as well as the results of our statistical analysis, tell us that we must go beyond chapter and consider the pedagogical and cognitive properties of the problems that we select. We found that problems which ask students to perform different procedural tasks (e.g. making a flow chart of how you would solve a problem) are important to distinguish experts from novices. Lastly, we found that "easy" problems—as determined by a typical introductory physics textbook author—did a better job of discriminating experts from novices.

Because of the different variability explained by the 5 and 10 problem analyses, we also conclude that there is some set size dependence on the ability to distinguish experts from novices. However, analyzing the sub-graphs of our sorters for a particular subset is not the same as giving our sorters fewer problems. Depending on the subset of problems chosen, it is possible for all of the problems to be in a single category or for each problem to be in a unique category. Based on anecdotal evidence from discussions with our sorters, this kind of extreme behavior would be highly unlikely. We therefore cannot comment further on optimal group size at this time.

Finally, a limitation of the current analysis is that only 42.7% of the variability in the dataset is explained by the problem variables that we have included in this analysis. We believe that the outstanding variability in our dataset is due to some as yet unknown internal cognitive processes. This means that there are other latent variables–perhaps linked to the surface features in each problem–which we must consider in order to explore this relationship more fully. Understanding this outstanding variability will be the key to predicting an instrument's ability to discriminate experts from novices.

ACKNOWLEDGMENTS

The authors would like to thank the MSU physics faculty and the introductory physics classes in the Fall 2010/Spring 2011 semester for volunteering to sort problems for our study. One of the authors (SW) would also like to thank R. Teodorescu for her help in understanding the various nuances of the TIPP, as well as her assistance in constructing problems which would be diverse along both dimensions of the TIPP. Also, SW would like to thank Brian O'Shea for his help in getting an account on the High Performance Computer Cluster in order to run this analysis. One of the authors (GK) would also like to thank the group of David Pritchard at MIT for its hospitality during his sabbatical.

REFERENCES

1. M. T. H. Chi, P. J. Feltovich, and R. Glaser, *Cognitive Science* **5**, 121 – 152 (1981), ISSN 0364-0213.
2. S. F. Wolf, D. P. Dougherty, and G. Kortemeyer, *Phys. Rev. ST Phys. Educ. Res.* **8**, 010124 (2012).
3. R. E. Scherr, *American Journal of Physics* **75**, 272–280 (2007).
4. B. W. Frank, S. E. Kanim, and L. S. Gomez, *Phys. Rev. ST Phys. Educ. Res.* **4**, 020102 (2008).
5. A. Mason, and C. Singh, *Phys. Rev. ST Phys. Educ. Res.* **7**, 020110 (2011).
6. J. S. Walker, *Physics*, Pearson Education, Inc., Upper Saddle River, New Jersey 07458, 2004, second edn., ISBN 0-13-101416-1.
7. R. Teodorescu, C. Bennhold, and G. Feldman, "Enhancing Cognitive Development through Physics Problem Solving: A Taxonomy of Introductory Physics Problems," in *Physics Education Research Conference 2008*, Edmonton, Canada, 2008, vol. 1064 of *PER Conference*, pp. 203–206.
8. C. Singh, *American Journal of Physics* **77**, 73–80 (2009).
9. H. Hotelling, *The Annals of Mathematical Statistics* **2**, pp. 360–378 (1931), ISSN 00034851.
10. L. Baringhaus, and C. Franz, *Journal of Multivariate Analysis* **88**, 190 – 206 (2004), ISSN 0047-259X.
11. B. A. Wiggins, *Applied and Environmental Microbiology* **62**, 3997–4002 (1996).
12. H. Hotelling, *Biometrika* **28**, 321–377 (1936).