

Thinking About Music: Novice and Expert Inductive Reasoning

Liz Baraff (liz_b@mit.edu)

MIT Brain and Cognitive Science Department, 77 Massachusetts Ave; Bldg. NE20-388
Cambridge, MA 02139 USA

John D. Coley (j.coley@neu.edu)

Northeastern University Department of Psychology, 360 Huntington Ave, 125 NI Hall
Boston, MA 02115 USA

Abstract

Recent research (e.g. López, Atran, Coley, Medin & Smith, 1997; Proffitt, Coley & Medin, 2000; Shafto & Coley, in press) has revealed striking expert-novice differences in category-based induction in the domain of folk biology. In this paper we examine the generality of those findings by investigating expert-novice differences in category-based induction in the domain of music. Experiment 1 revealed that experts and novices showed extremely high agreement in terms of how they sorted the names of 24 musical composers into groups. Experiment 2 employed a standard strength-of-argument rating task to assess the degree to which measures of taxonomic distance derived from Experiment 1 predicted category-based inferences. Results were precisely as previously reported for folk biology; novices demonstrated effects of both premise-conclusion *similarity* and premise *diversity*, where experts showed *similarity* but not *diversity*. Experiment 3 replicated Experiment 2 except that experts and novices both rated argument strength under speeded conditions. Under cognitive load, premise-conclusion *similarity* persisted for both experts and novices. In contrast, under cognitive load novice premise *diversity* disappeared, whereas for experts diversity was evident only under cognitive load. These results suggest that patterns of reasoning previously reported for folk biological induction may be more generally applicable. They also suggest important processing differences between experts and novices.

Introduction

In even the most mundane day, we are presented with numerous situations requiring inductive inferences. Often these are based on relations among categories. For example, you know you are very allergic to peanuts and cashews but don't know if you are allergic to any other nuts. If someone offers you brownies with walnuts in them, you have to weigh the likelihood that walnuts also have the property "causes allergic reaction" before accepting. These inductive inferences do not necessarily have right or wrong answers, but are instead perceived as relatively strong or weak.

Besides the regularity and importance of inductions, they also tell us something about our underlying category structure. One model used to describe this is the Similarity-coverage Model (SCM), developed by Osherson et al. (1990). Osherson found that single premise arguments were stronger when the premise categories were more similar to the conclusion category. However, dual premise

arguments with dissimilar premise categories were rated stronger than dual premise arguments with similar premise categories. Osherson et al. proposed a similarity-coverage model (SCM) to explain the results. Two important phenomena are:

Premise-Conclusion Similarity: arguments whose premises are very similar to the conclusions are rated stronger than arguments whose premises are different from the conclusion.

Premise Diversity: arguments whose premises are very different from one another, and thus cover more of the category space, are rated stronger than arguments whose premises are similar to one another and thus have less coverage.

While the similarity-coverage model describes typical, American, undergraduate, participant responses, more recent research suggests that experts may reason differently from these novices. Proffitt, Coley, and Medin (2000) tested tree experts on inductive reasoning tasks where arguments had to be rated in terms of the likelihood of certain trees getting disease 'x' when other trees showed symptoms. Their results show that instead of using the diversity/coverage of the premises as the SCM predicts, experts are instead influenced strongly by causal-ecological factors. This suggests that experts are using detailed specific knowledge rather than general taxonomic relations to guide induction. (See also Lopez et al., 1997; Shafto & Coley, in press.)

In response to this research, we were interested in answering several important questions. Previous research has focused on testing inductive reasoning in the domain of biology, where rich taxonomic structures are already available and salient perceptual features help guide these taxonomies. Animals that look similar are also very likely to be taxonomically/biologically similar; thus, novice and expert categories that may be based on different member features are categorized similarly because the features are highly correlated. Our first task is to examine categories in a more abstract domain by testing how experts and novices classify musical composers.

The second task will use the category structure established from Experiment one to derive inductive reasoning questions. We can then test predictions derived from the SCM in the new domain of music. If consistent

with the SCM, novices should use similarity and coverage to rate strength of arguments as they do in biological inductive reasoning tasks. Our second experiment also tests expert inductions. If our findings are consistent with previous research, music experts should use similarity to compute argument strength, but may use context dependent knowledge to compute other arguments, forgoing diversity.

Our third study seeks to test a computational prediction that Osherson's model indirectly makes. According to the SCM of Osherson et al (1990), the similarity phenomenon involves computing similarity between premise and conclusion categories, whereas the diversity phenomenon involves accessing members of a general conclusion category and computing similarity between premise categories and sampled members of the conclusion category. Therefore, computing coverage (the proposed mechanism for diversity) is more complex and involves more steps than computing similarity. As such, if the SCM is an appropriate model of novice reasoning, then for novices a cognitive load (e.g., performing a speeded reasoning task) should lead to a decrement in the diversity phenomenon while leaving the similarity phenomenon intact.

For experts, the predictions are more complex. When evaluating category-based arguments, experts appear to favor specific knowledge rendered relevant by relations among categories and properties over general taxonomic relations among categories; thus, the SCM fails to predict expert responses to diversity items. One possible account of this failure is that experts retain a general scheme of taxonomic relations among concepts in their domain of expertise, but also acquire a rich network of specific thematic relations that augment and potentially override taxonomic relations in guiding inference. If accessing this expert knowledge is relatively slow and effortful, a cognitive load may block its application. If so, expert performance under cognitive load may approach novice performance without such a load, and may conform more closely to the predictions of the SCM.

Experiment 1

The first experiment is designed to test music expert and novice category structure based on sorting task similarity. Previous research has show high levels of agreement between expert and novice sorting in the domain of biology. (Lopez et al, 1997; Shafto & Coley, in press.) By examining participants' category sorts in a more abstract domain like music, with fewer feature correlations than biological domains, we may reveal more striking expert novice differences.

Method

Participants: Twenty-four participants were run in total. Of the twenty-four twelve were novices and twelve were experts. The novice group included participants varying in age from 18 to 45. At the end of the experiment, all novice participants were asked if they had any extensive music training and/or any courses in ethnomusicology so that they

could be ruled out of the study as too knowledgeable. No participant was considered too knowledgeable, and thus data from all twelve participants was retained. The twelve experts were musicians and composers in the greater Boston area including professors at Northeastern University, MIT, Boston University, Berklee School of Music, and New England Conservatory, and graduate students studying music at these institutions.

Materials/Design: There were twenty-five index cards, each with a different composer's full name typed using normal font. Names were selected after a pilot study indicated which composers might be better known by novices to avoid false sorting. Experts and novices were compared.

Procedure: Participants were given category cards and asked to set aside any composers that they did not feel familiar with or comfortable categorizing. After these cards were recorded and removed, participants were asked to sort based on music composition style similarity. Participants were allowed to make as many groups as they wanted in their initial sort, but were told that they would have the opportunity to further define each group. Participants continued to subdivide groups based on composition similarity until they felt they could no longer use similarity as a differentiator. Cards were then regrouped into the participant's initial sort and participants were asked to combine any groups that they found more similar based on composition style.

Results

Expert and novice sorts were analyzed by using a software program designed to output a 24X24 symmetric pair-wise distance matrix that calculates each value by taking the number of steps required to get to the lowest level shared category of each composer. For example, if two composers were continually grouped together through the very smallest sub grouping, their distance would be 1. If two composers were never grouped together, and the participant regrouped 7 times, their distance would be 7. Because each participant used a different total number of groupings, each participant's matrix was standardized.

The values from each matrix were then analyzed using the Cultural Consensus Model, (for similar applications, see Lopez et al. 1997). The Cultural Consensus Model examines agreement between participants via factor analysis. A single factor solution can be taken as evidence of a single underlying body of knowledge. Factor analysis of an inter-participant agreement matrix revealed high agreement among experts and novices. The first factor accounted for 72% of the variance and was 9.6 times greater than the second factor. All participants had a positive first factor score. This suggests a single, same underlying category structure for experts and novices.

We also performed a multi-dimensional scaling analysis of mean pair-wise distance among composers. A 2-dimensional solution accounted for over 97% of the

variance. As can be seen in figure one, composers clustered into four groups.

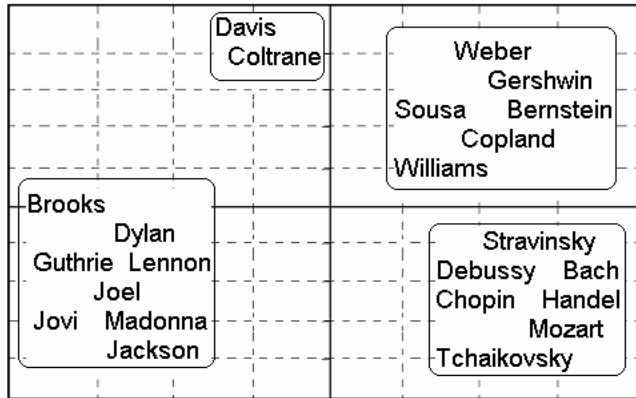


Figure 1: The 2-dimensional model of expert and novice sorting results.

Discussion

The results from Experiment one show that there are few differences in category structure between novices and experts. There are two possible explanations for this outcome. Either the differing salient properties that novices and experts are using to complete their sorts are aligned, or experts and novices use the same properties to differentiate groups in the domain of music. As in biology, agreement between experts and novices is remarkably high.

Experiment 2

The second task will use the category structure established from Experiment one to derive inductive reasoning questions. We can then test predictions derived from the SCM in the new domain of music. If novices' responses are consistent with the OCM then they should show the phenomenon of premise-conclusion similarity and premise diversity as discussed above. If expert responses replicate previous research, then they should use similarity but not diversity.

Method

Participants: 17 Participants were run in the second experiment. 12 participants were classified as novice preceding the experiment, and nine expert participants. Novice participants included Northeastern University Undergraduates completing the experiment for course credit. Experts were musicians and composers around the community including professors at Northeastern University, MIT, Boston University, Berklee School of Music, and New England Conservatory, and graduate students studying music at one of these institutions.

Material/Designs: The instructions, test trials, and questions in Experiment 2 were presented using the PsyScope computer program. There were twenty-eight

questions total randomly presented during the computer portion of the experiment.

We explored the three phenomena, similarity, diversity specific, and diversity global by creating 6 sets of questions. (See Table 1.)

Similarity: Two sets of questions were created to assess similarity, a strong similarity set and a weak similarity set. The strong similarity set consisted of arguments whose premises were always taken from the same category. The conclusions for strong similarity questions were also taken from the same category as the premises. In the weak similarity set premises were taken from the same category, however conclusions were taken from different categories from the premises.

Diversity Global: Additionally, one set of questions tested strong diversity global questions, and one set tested weak diversity global questions. The strong set consisted of diversity global questions where premises were taken from different categories. In the weak set, diversity global questions were constructed with premises taken from the same category. For both strong and weak diversity global sets conclusion categories were always general.

Diversity Specific: The last two sets of questions consisted of strong diversity specific questions and weak diversity specific questions. Strong sets were constructed so that premises were taken from different categories, and conclusions were taken from a third and different category from the premises. Weak sets were constructed so that premises were taken from the same category, but the conclusion was taken from a different category than the premise category.

	Strong	Weak
Similarity 8 questions total	<u>Mozart, Bach</u> Beethoven	<u>Lennon, Joel</u> Beethoven
Diversity: Global 16 questions total	<u>Handel, Madonna</u> All Composers	<u>Handel, Chopin</u> All Composers
Diversity: Specific 8 questions total	<u>Bach, Dylan</u> Miles Davis	<u>Bach, Mozart</u> Miles Davis

Table 1: Shows types of questions asked in Experiment two.

Procedure: PARTICIPANTS were tested individually and told to take their time and think about each question carefully before answering. The computer presented the question first in red for 15 seconds where participant responses were not recorded. After the 15-second interval, the question turned green indicating that the participant could now input their response. Each argument could be rated on a scale from 1-weak argument to 7-strong argument.

Results

Novice and expert results were analyzed using two-sampled unequal variance, 2-tailed t-tests by participant and question type. For each of the three phenomena mean strength ratings were compared for strong vs. weak

arguments via t-test. Results for both novices and experts are presented in Figure 2 as difference scores (mean strong rating minus mean weak rating) for each phenomenon.

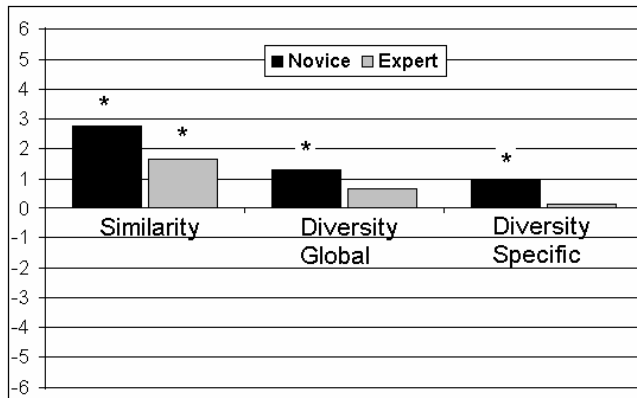


Figure 2: Novice and expert difference score in unsped condition.

Similarity: Novices rated similarity questions significantly higher than non-similar questions, ($p < .01$). Experts also rated similarity questions significantly higher than non-similar questions, ($p < .01$).

Diversity Specific: Novices also used diversity, rating diversity specific questions higher than non-diversity specific questions, ($p < .05$). However, experts did not use diversity; specific questions were not higher than non-diversity specific questions, ($p = .65$).

Diversity Global: Novices also rated diversity global questions significantly higher than non-diversity global questions, ($p < .05$). However, experts again did not use diversity and rated strong global diverse question equally as high as weak global diverse questions ($p = .28$).

Discussion

Just as research in biological domains suggest, when reasoning about music, novices use the degree of similarity and coverage to rate arguments' strength. Expert responses also replicate previous findings by using similarity to determine argument strength, but appear to be ignoring diversity/coverage just as Proffitt et al. (2000) suggest. Instead, experts are using some information not expressed in their category model to determine diversity argument strength. Thus, the results from Experiment 2 replicate the results of previous research in other domains suggesting that there may be domain general properties of expert knowledge.

Experiment 3

Our third study seeks to test the SCM's computational prediction that computing similarity is less cognitively taxing than computing diversity. As such, if the SCM is an appropriate model of novice reasoning, then for novices the cognitive load should lead to a decrement in the diversity phenomenon while leaving the similarity phenomenon

intact. For experts, the predictions are more complex since the SCM fails to predict expert responses to diversity items in a slow task. It is possible that accessing expert knowledge is relatively slow and effortful; thus, a cognitive load may block its application. If so, expert performance under cognitive load may approach novice performance without such a load, and may conform more closely to the predictions of the SCM.

Methods

Participants: 12 participants were classified as novices preceding the experiment, and six as experts. Novice participants included Northeastern University Undergraduates completing the experiment for credit. Experts were musicians and composers around the community including professors at Northeastern University, MIT, Boston University, Berklee School of Music, and New England Conservatory, and graduate students studying music.

Materials/Design: The materials and design of Experiment 3 are exactly the same as Experiment 2.

Procedure: The speeded task was identical to the slow task with one exception. Instead of a forced 15-second waiting time between question presentation and response, the participant had a short, 3-second interval before being able to answer the question. Here, the question was first presented in red. After three seconds, the question turned green indicating that responses could be entered. The participants were also encouraged to answer as quickly as possible without sacrificing accuracy for speed. In all cases, the task took less than 30-minutes to complete.

Results

Cognitive Load: Novice and expert results were analyzed using two-sampled unequal variance, 2-tailed t-tests by participant and question type. For each of the three phenomena mean strength ratings were compared for strong vs. weak arguments via t-test. Results for both novices and experts are presented in Figure 3 as difference scores (mean strong rating minus mean weak rating) for each phenomenon.

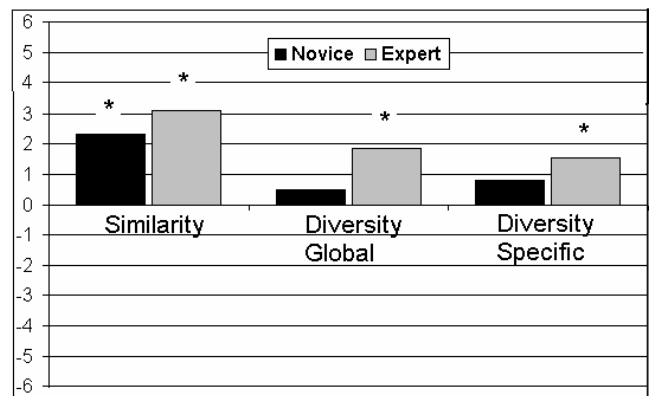


Figure 3: Novice and expert difference scores in speeded condition.

Similarity: Novices rated strong similarity questions significantly higher than weak similar questions, ($p < .01$). Experts also rated strong similarity questions significantly higher than weakly similar questions, ($p < .01$).

Diversity Specific: Novices did not use diversity in the speeded task, rating strong diversity specific questions no higher than weak diversity specific questions, ($p = .06$). Interestingly, experts did use diversity in the speeded task and rated strong diversity specific questions significantly higher than weak diversity specific questions, ($p < .01$).

Diversity Global: Similarly, strong diversity global question were not rated significantly higher than weak diversity global questions for novices, ($p = .26$). Again, however, experts did rate strong global diversity questions significantly higher than weak ones ($p = .01$).

Comparison of Experiments 2 & 3: Expert and Novice response times were compared in Experiment two and three. A paired T-test revealed that expert response times were significantly longer than novice response times ($p < .01$) in Experiment two. There were no differences between expert and novice response times in Experiment 3, ($p = .56$).

To directly examine the impact of cognitive load, difference scores from the unsped task were compared to difference scores from the speeded for both novices and experts for each phenomenon. A t-test revealed that novices can use the similarity equally as well in fast and slow conditions ($p = .34$). Although novices did not show diversity in the speeded task, there is no significant difference between their performance in the speeded and unsped conditions; a t-test revealed no significant difference between their diversity global scores ($p = .31$) in fast vs. slow conditions, or their diversity specific scores ($p = .83$) in fast vs. slow conditions. In contrast, a T-test comparing fast and slow experts revealed significant change of similarity ($p < .05$), diversity specific ($p < .01$) and near significant changes in diversity global ($p = .07$). (See Figure 4)

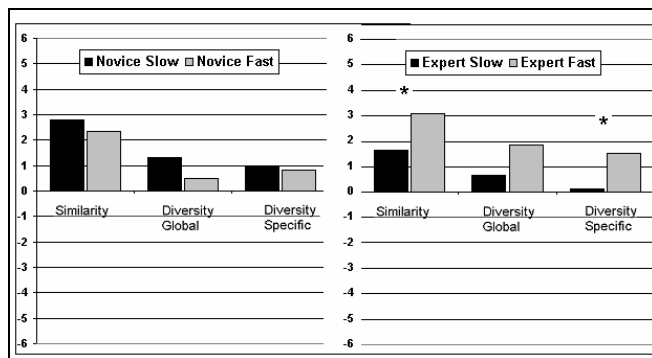


Figure 4: Novice and expert difference scores in fast and slow trials.

Discussion

While speed didn't seem to affect similarity responses for either novices or experts, speeding the task did greatly affect

diversity. Novices rated similarity arguments significantly higher than non-similar arguments in both fast and slow conditions, but rated diversity specific and global significantly higher than non-diversity specific and global arguments in only the *slow* condition. However, experts rated similarity significantly higher than non-similar questions in both fast and slow conditions, but rated diversity specific and global significantly higher than non-diversity specific and global questions in only the *fast* condition. Moreover the comparison of the response time in Experiments two and three suggest that experts may have been using some computationally more taxing reasoning in their judgments for Experiment two only. Comparison of differences scores from Experiments two and three suggest that the speeded condition had a greater impact on experts than on novices.

General Discussion

Overall, the pattern of results is clear. Without a time constraint, novice responses were as predicted by the SCM for both premise-conclusion similarity and premise diversity, whereas experts failed to show any effect of diversity. This clearly fits with previous research in folk biology; expertise leads to a reduction in diversity-based reasoning, presumably because of the availability and salience of numerous orthogonal relations. While it is not clear specifically what type of reasoning music experts are abandoning diversity for, it is likely that their responses are contingent on the specific people and properties addressed in each question. For example, one expert explained informally that Mozart and Bon Jovi--a diverse premise pair predicted to have relatively high coverage based on sorting data--were actually quite similar in that both use strong beats in their composition style. Likewise, this expert pointed out that Mozart and Debussy--both from the classical group and therefore predicted to have relatively low coverage--were actually quite different as Debussy's use of free-from rhythm contrasts sharply with Mozart's use of a strong beat. Novices, lacking such detailed specific knowledge, rely on the taxonomic relations revealed by their sorting. One bit of evidence in support of this explanation is the fact that experts took significantly longer than novices to respond in the unsped condition, suggesting the use of rich, context dependent relations that are more cognitively taxing than novice taxonomic relations. Expert and novice response times did not differ in the speeded condition.

Under the cognitive load of making speeded judgments, premise-conclusion similarity persisted for both experts and novices, suggesting that computing similarity among given composers involved a relatively quick computation impervious to this level of cognitive load. In contrast, speeding judgments changed both experts' and novices' responses to premise-diversity items. For novices, diversity effects disappeared, supporting the prediction of the SCM that computing coverage is more cognitively taxing than computing similarity. For experts, diversity effects are

apparent under cognitive load only. This suggests that the relations used by experts in unspeeded reasoning--which lead experts to show no effects of premise diversity--may be difficult to access under cognitive load. It also suggests that the taxonomic relations that underlie the premise-diversity phenomenon are both present and highly accessible to experts who, unlike novices, appear able to compute coverage under speeded conditions.

In sum, these results suggest qualitative differences in the process by which experts and novices evaluate the likelihood of inductive arguments. Experts' default approach to the task utilizes their rich, context-dependent and relatively deliberative knowledge base. Under cognitive load, experts can no longer access this knowledge, but can utilize the taxonomic knowledge evident in Experiment 1. Thus, although not preferred, taxonomic relations are highly accessible to experts. In contrast, novices--who by definition lack rich context-dependent domain knowledge--utilized taxonomic relations as a default approach. Experiment 1 showed that these relations are similar to those of experts, but Experiment 3 suggests that they are not as readily available under cognitive load. A direct comparison of results in speeded and unspeeded conditions supports this view. Cognitive load induced a qualitative change in experts' approach to induction, as indicated by the reliable differences between the two conditions. In contrast, no such change was evident for novices. Although sufficient to eliminate diversity effects within Experiment 3, cognitive load did not lead to reliable condition effects for novices. Our results suggest that patterns of reasoning previously reported for folk biological induction may be more generally applicable. They also suggest important processing differences between experts and novices.

Acknowledgments

We would like to thank the Northeastern University Categorization and Reasoning Lab, Fei Xu and Rhea Eskew for their support and input, and especially Rachel Kolter for her help in data collection and constructive comments.

References

- Lopez, A., Atran, S., Coley, J.D., Medin, D. & Smith, E.E. (1997) The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology*, 32, 251-295.
- Osherson, D.N., Smith, E.E., Wilkie, O., Lopez, A. & Shafir, E. (1990). Category-based induction. *Psychology Review*, 97(2), 185-200.
- Proffitt, J.B., Coley, J.D. & Medin, D.L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 26(4), 811-828.
- Shafto, P. & Coley, J.D. (In press). Development of categorization and reasoning in the natural world: Novices to experts, naïve similarity to ecological

knowledge. *Journal of Experimental Psychology: Learning, Memory & Cognition*.