The transfer of logically general scientific reasoning skills

Anthony M. Harrison (anh23@pitt.edu)
Christian D. Schunn (schunn@pitt.edu)
Department of Psychology, University of Pittsburgh
3939 O'Hara St
Pittsburgh, PA 15260 USA

Abstract

Extending the paradigm introduced by Schraagen (1993), two near-expert groups and novices completed two scientific discovery tasks, one from each of the experts' domains. In this way, both groups designed simulated experiments from within and outside of their domain. The role of domain familiarity on the application of general scientific reasoning skills is explored by contrasting the performance of the experts in their domain to that in the unfamiliar domain. Results indicate that at the graduate level, near-experts are able to apply general scientific reasoning skills across dissimilar domains, while novices still have difficulty with the transfer.

Introduction

How common are scientific reasoning skills across different domains of practice? Schraagen (1993) as well as Schunn and Anderson (1999) address this issue in their experiments. Both studies relied on the same basic paradigm: two groups of expert researchers and one group of novices were asked to design and conduct a series of experiments in a scientific discovery task. One of the expert groups was familiar with the domain that the task was drawn from (e.g. cognitive psychologists working on a memory experiment), whereas the other was less familiar with the domain, but still from the same general field (e.g. social psychology). The scientific reasoning skills (e.g. experimental design, hypothesis generation, and data evaluation) exhibited by both expert groups were similar and mapped cleanly onto those mentioned in the literature (e.g. Klahr & Dunbar, 1988; Dunbar, 1993; Chinn & Malhotra, 1999). The novices showed a similar pattern of failures as those found by other researchers focusing on non-scientists (e.g. Kuhn, Schauble & Garcia-Mila, 1992; Klahr, Fay & Dunbar, 1993; Zajchowski & Martin, 1993).

Almost all studies that have focused on non-scientists have found remarkably poor performance (see Detterman, 1992 for a review) on scientific reasoning skills. Fortunately, those that have studied practicing scientists in their own domain have found just the opposite (e.g. Dunbar, 1997). Why is it that expert scientists do so well outside of their domain of expertise, yet novices do so poorly? Schunn and Anderson (1999), as well as Schraagen (1993), report that the differences found between the practicing scientists and the novices could not be accounted for by general reasoning ability differences. This leaves two alternative explanations: context of the problem influencing the transfer of the skills, and scientific training itself.

While the studies conducted by Schraagen (1993) and Schunn and Anderson (1999) seem to point towards the influence of scientific training, there is a problem with interpreting it in that way. Both of these studies utilized highly similar groups of experts (all were experimental psychologists of some sort). The similarity of the domains of expertise might be confounding the influence of context on transfer. While the experiments were designed such that the task would be unfamiliar to one of the expert groups, it is likely that they were familiar enough to provide sufficient context cues to trigger the use of the appropriate scientific strategies. The novices, however, would not have had the cues to signal which strategies would be appropriate. If the studies had utilized more dissimilar experts this would not have been an issue. As it is, the transfer context remains a confounding factor.

Voss et al. (1986) provide some support for the transfer hypothesis. Their studies looking at the problem solving of novices and dissimilar experts found that the quality of the reasoning was dependent upon the match of the task to the expert's domain. They report chemists' reasoning being roughly equivalent to that of the novices when attempting to solve political science problems, while the political scientists exhibited a higher quality of reasoning on the same tasks. However, while the Schraagen and Schunn & Anderson studies utilized science-based domains that were overly similar, the Voss, et al. study utilized a non-science-based domain. This makes comparisons across the studies problematic.

The difficulty of transferring skills from one context to another has long been a recognized problem (e.g. Thorndike & Woodworth, 1901; Singley & Anderson, 1989; Detterman, 1992). Singley and Anderson propose that transfer can only occur between two situations for identical elements. This transfer then depends on how the elements were encoded. If researchers only use scientific reasoning skills when faced with problems in their domain, it is possible that they will be coded in a manner that is specific to that context. It would therefore be unlikely that they would use the strategies when those context cues were absent. So in this situation, the

superficial elements of the unfamiliar task mask the relevance of using scientific skills. Unless there is a more abstract understanding of these skills (more general encoding), they will fail to be used in other scientific reasoning contexts.

Much like the experiments of Schraagen (1993) and Schunn and Anderson (1999), this study was designed to explore the differences in scientific reasoning in novices, task experts (at experimentation in general) and domain experts (in the problem domain). The core difference is that instead of using a single task, there are two isomorphic discovery tasks from different scientific domains. Each of the expert groups is a domain expert in one of the two tasks, making each group task experts for both and domain experts for only one. This design allows us to look specifically at the influence of domain familiarity on the transfer of scientific reasoning. Should domain familiarity play a key role in the transfer of the scientific reasoning skills, we would expect the skills to only manifest themselves when the experts are within their natural domain. When working in the unfamiliar domain, their performance would be more like that of the novices. If, however, they are able to recognize the deeper scientific structure of the unfamiliar problem, then their performance should be better than that of the novices and qualitatively equivalent to that seen in their natural domain.

Methods

Participants were recruited from two major universities: 33 undergraduates, 16 from one university, and 17 from the other. The graduate samples were each drawn from a different university (due to enrollment). 11 biology graduate students and 17 industrial/organizational psychology students were recruited, all had completed at least two years of study. Participants were paid for their participation.

Participants were asked to complete two isomorphic experiment-based exploration tasks: they were to determine how each of six task relevant independent variables (IV) affects the outcome of the dependent variable (DV). For example, the biology task had participants design experiments to determine how each variable (water temperature, turbidity, dissolved oxygen, pH, fecal and phosphorus contents) influenced the growth of a certain bacteria that was responsible for the development of open sores on fish. The industrial psychology task had participants design experiments to determine the role of manager characteristics (e.g. technical, critical reasoning, writing skills) on the objectivity of employee appraisals.

Experiments were designed and conducted in a computer simulated laboratory where participants were able to manipulate each IV in question; run and view experiments; take notes on hypotheses, experiments, and outcomes; as well as assign conclusions as to the influences of the IVs on the DV. No data analysis or graphing tools were provided. Each task was self-paced, allowing participants to conduct as many experiments as they needed in order to draw their conclusions. The simulated experiment lab was built with Java™ allowing the recording of all user actions for playback[1].

The task domains were selected based on graduate enrollment in the domains across two universities. Experts in the domains (instructors and researchers) were recruited during the design of the tasks to maximize the external validity of the tasks. The tasks, from microbiology and industrial/organization psychology, were based on experiments found in the literature. The IVs' qualitative effects on the DVs were maintained for 4 of the 6 variables. Two IVs from each task were modified so that they would produce anomalous results. One was anomalous from a theoretical perspective (TA), which was merely an inversion of the qualitative trend. The second was data anomalous (DA) in that a 20% subset of the variable's range produced extreme values. The anomalous variables were added to test the sensitivity of the participants to data anomalies (e.g. Chinn & Brewer, 1993).

Results

The critical comparisons in this study are two orthogonal comparisons: the experts vs. novices (graduates and undergraduates), as well as between the two groups of experts. The expert analyses are between the experts within their domain and outside their domain (domain and task experts respectively). The In-Domain group consists of the biology graduates working on the biology task and the psychology graduates solving the psychology task. The Out-Domain group consists of the same participants, merely performing the opposite task. Unless otherwise noted, all tests are repeated measure ANOVAs with the two tasks as the repeated factor. Specific statistical measures are reported in table 1; significance is assumed at $p<0.05$. Task presentation was counter-balanced, and there were no significant effects of task order on any of the reported measures.

Experimental Design Measures

Participants spent on average 43 minutes on each of the two tasks. While the graduate students spent a little longer on each task (approx. 47 minutes) than the novices (approx. 39 minutes), the differences were nonsignificant. Likewise, the number of experiments designed and conducted in each task did not differ significantly either between the novices and the experts (37 and 47

[1] The experiment can be downloaded from the author's website http://simon.lrdc.pitt.edu/~harrison/

respectively) or within the In-Domain and Out-Domain groups (48 and 45 respectively).

The next design measure considered the breadth of the experimental space that the participants searched (Klahr & Dunbar, 1988). Each IV that they could manipulate had a fixed range, which were divided into five equally-sized bins. For each unique experiment that manipulated a given variable, the total number of bins visited was computed. The more complete the variable range covered, the more informative the results will be with respect to that variable on the whole.

As expected, the novices covered a significantly smaller range than the graduates (see figure 1). However, there were no differences between the In-Domain experts (e.g. biology graduate students performing the biology task, and vice versa) and the Out-Domain experts (e.g. psychology students performing the biology task).

Looking at the breadth of search for the two anomalous variables yielded similar findings. The novices searched a much narrower space for the theory anomalous and data anomalous variables. Likewise, there were no significant differences between the domain and task experts.

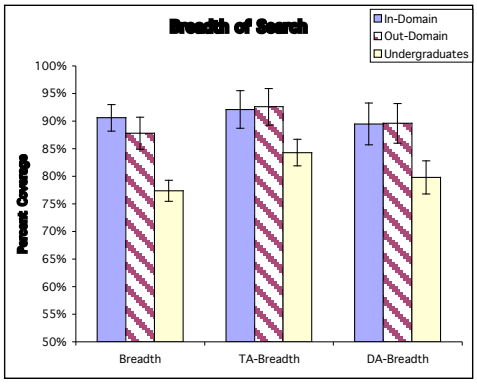

Figure 1. Breadth of experimental space search for all, theory anomalous (TA), and data anomalous (DA) variables.

Another design measure looked at the conservativeness of the experimental designs. With the lack of any data analysis tools, maximizing the interpretability of each simulated experiment was very important. Participants conducted experiments one at a time. The only way to draw conclusions was to compare outcomes of successive experiments. If the comparisons were confounded (multiple IV manipulations), accurate conclusions would be very difficult to draw. Two versions of a VOTAT (vary one thing at a time) score were computed for each task (Tschirgi, 1980). The local VOTAT was computed by averaging the number of variables manipulated in one experiment when compared to the immediately previous experiment. The global VOTAT was computed between the current experiment and the most similar previous experiment (regardless of when it occurred). Since there were no significant differences between the two measures across groups and tasks, the two were combined into an average composite. The novices consistently manipulated multiple variables per experiment, averaging 1.42 changes per experiment, which was significantly more than the graduate sample's 1.23. There were no differences between the experts within or outside their domains.

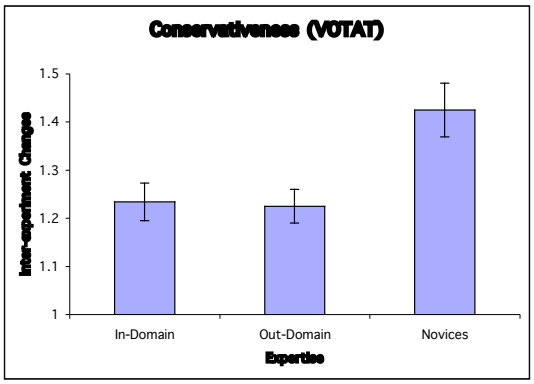

Figure 2. Average number of variables manipulated per experiment. One change at a time yields optimal interpretability in this scenario.

Note-taking

It is hard to argue against the importance of meta-processing skills in scientific reasoning. For this study, we chose to look at the note taking behavior of the participants. This is analogous to practice of scientists keeping a detailed lab notebook (Dunbar, 1997). The first measure is merely one of length, how much note taking is going on. Novices took very few notes (if any), which is in stark contrast to the experts who took significantly longer notes. For the experts, there were no differences in note-taking length when in or out of their natural domain.

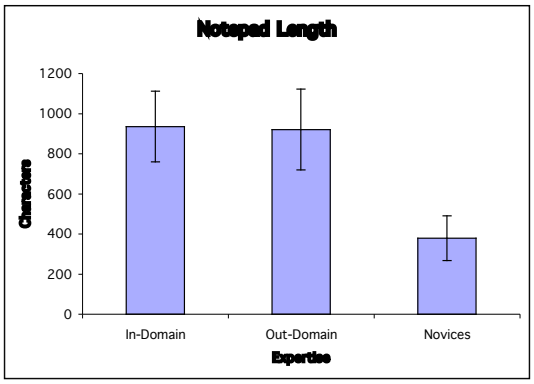

Figure 3. Length of notes (in characters) taken during experiment. Significant difference between novices and experts, none between the experts.

Knowing how much note taking was occurring lead us next to ask how they were utilized. A simple three-category scheme was developed a priori. Notes could be categorized as non-existent (including uninterpretable and

irrelevant notes), effects tracking (documenting variable values and experiment outcomes, effectively duplicating the provided experiment log), or hypothesis tracking (documenting hypotheses, suspected relationships, etc.). As can be seen in figure 4, all groups did an equal amount of effects tracking. The significant differences were in the amount of hypothesis tracking that the experts did.

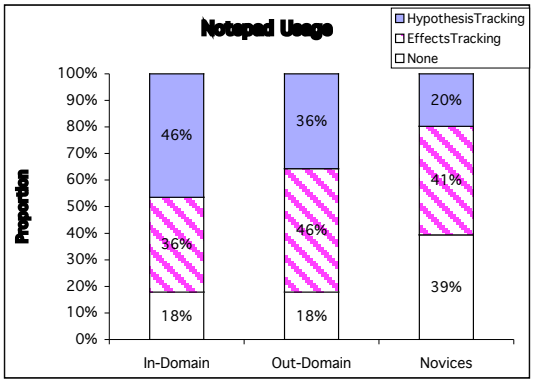

Figure 4. Type of notepad usage. Significant difference between novices and experts, $\chi^2(2,N=61)=5.9$, $p<0.05$. No significant differences between expert groups.

Data Interpretation

The final set of measures where designed to assess the participants' ability to interpret the data and draw conclusions regarding the relationships between the independent variables and the outcome measure. Participants were asked to write out their conclusions for each variable including the qualitative trend, critical values, magnitude, and any "strange" properties. One point was awarded for each property that was correct. Averaging across each of the variables yielded an overall interpretation accuracy score. Making it conditional on the breadth of the experiment space that was searched further refined each variable's score. For instance, any conclusions about critical values would be erroneous if they had not explored the variable range within which the values occurred. The left most graphs in figure 5 show both the raw and the conditional overall interpretation accuracy scores. Once more the novices scored lower than the experts with no significant differences between experts in or out of their domains. The differences between the raw and conditional scores were nonsignificant; participants did not appear to be drawing conclusions beyond what was possible given the data collected.

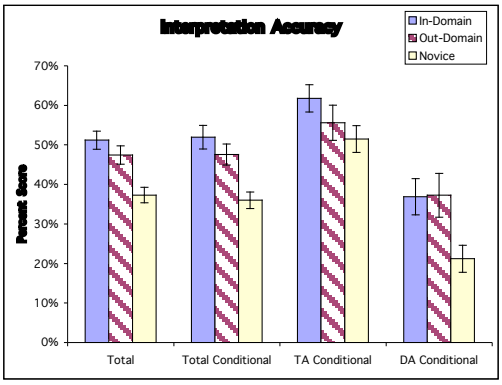

Figure 5. Interpretation accuracy scores, conditional on breadth of search.

The conditional accuracy scores for the anomalous variables were also examined (right half of figure 5). For the theoretically anomalous (TA) variables there were no differences among the three groups. The difference between In- and Out-Domain participants for the TA variable would likely be magnified if established domain experts had been used instead of the graduate students. There was a significant difference between novices and experts for the data anomalous (DA) variables, but this is directly attributable to the differences in the breadth of search for the data anomalous variables (see figure 1).

IQ Surrogates

Since this study had to be conducted across multiple universities, one of the first concerns was general IQ differences between the sampled populations. Using SAT and GRE scores as IQ surrogates, simple ANOVAs were computed. There were no significant differences within the novice undergraduate samples across the two universities, allowing the collapsing of the two groups. There were significant differences between the graduate samples (one from each university) and the novices. The post-hoc LSD showed no differences between the two graduate samples. Figure 6 shows the average combined SAT/GRE scores of the three groups. Further more, regression models were run on all the key dependent measures to test for IQ effects. None of the models approached even marginal significance: higher IQ participants did not perform better than the lower IQ participants. This data suggests that the IQ confound cannot account for the expertise effects found.

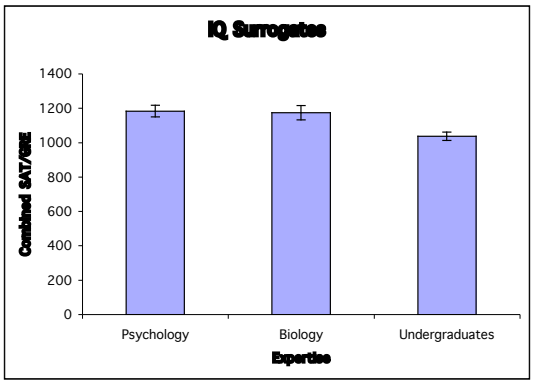

Figure 6. Combined SAT/GRE scores as IQ surrogates. No significant effects due to IQ.

Measure	Undergraduates v. Graduates (p-value)	In-Domain v. Out-Domain (p-value)
Breadth of search		
Overall	0.002	0.429
Theory	0.04	0.894
Data	0.02	0.987
VOTAT	0.02	0.86
Note length	0.001	0.93
Note usage†	0.05	0.875
Data interpretation		
Overall	0.00	0.234
Theory	0.41	0.301
Data	0.03	0.952
IQ surrogates††	0.000	0.34

Table 1. Significance of the planned comparisons. All measures are repeated measure ANOVAs with the tasks as the repeated factor, except for † Note usage (Chi-squared) and the †† IQ surrogate (standard ANOVA).

Discussion

Across the board expert groups performed significantly better than the novices, both inside and out of their domains. Additionally, there were no significant differences between the experts in or out of their natural domains. The findings mirror those of Schunn and Anderson (1999) and Schraagen (1993), even with the use of graduate students as opposed to established and practicing experts. Regardless of the domain, experts used the same general strategies in solving the discovery problems. Had the Out-Domain performance been significantly different from the In-Domain performance, bringing it closer to that of the novices, we could conclude that domain familiarity was influencing the transfer of the scientific reasoning skills.

The only major divergence from the former studies was the lack of a difference between the experts when it came time to draw conclusions from the data collected. While data interpretation per se has nothing to do with a particular domain, both Schraagen (1993) and Schunn and Anderson (1999) found qualitative differences between their task and domain experts. Both expert groups in this study performed equally poorly (but still significantly better than the novices). However, this could have been due to the fact that these tasks produced more data than those in other studies and lacked any analysis tools, such as graphs or tables (a common criticism leveled by the graduate students during the debriefing questionnaire).

Another difference in this study was the inclusion of anomalous variables (both theoretical and data based). Given that anomalies often draw the attention of scientists (Dunbar, 1997; Chinn & Brewer, 1993), it was interesting to see how sensitive the two expert groups were to them. One might expect that upon detection of an anomaly, that participants would search that section of the experimental space more thoroughly. Unfortunately, this is not seen for any of the expert groups (see figure 1). There are no reliably significant differences between the search patterns in terms of breadth or VOTAT (data not shown). This is not to say that the experts did not notice the anomalies; they just didn't exploit them in their experiments, in contrast to the findings of Dunbar (1997) and Tricket, et al (2001). This difference is likely due to the use of graduate students as opposed to established researchers.

Summary

Much like the studies of Schunn and Anderson (1999) and Schraagen (1993), we were interested in exploring the generality of scientific reasoning skills. Both studies concluded that while the quality of the reasoning was domain specific, the processes used were general. As was seen in these studies there are strong differences between the experimental skills exhibited by novices and the experts, or in this case, the developing experts. These differences cannot be attributed to a simple variable such as differences in intelligence. What's more is that the two expert groups behaved qualitatively the same whether they are working in their familiar domain or in a novel one. They were able to transfer the appropriate skills based on the deeper scientific structure and were not negatively influenced by the unfamiliar domain. The bidirectional transfer across a wider context difference is more evident in this study because of the utilization of genuinely dissimilar target domains, as opposed to the psychological domains used by Schunn & Anderson and Schraagen or the chemistry and (non-scientific) political-science domains used in the Voss, et al (1986) studies.

Acknowledgments

I would like to thank Lelyn Saner and Sasha Palmquist for their comments on earlier drafts of this work.

References

Chinn, C., & Brewer, W. (1993). Factors that influence how people respond to anomalous data. Proceedings of the 15th Annual Conference of the Cognitive Science Society, USA, 318-323

Chinn, C. A., & Malhotra, B. A. (1999). Epistemologically authentic scientific reasoning. In Crowley, K., Schunn, C.D., & Okada, T (Eds.), Designing for Science:Implications from Professional, Instructional, and Everyday Science. Mahwah, NJ: Erlbaum.

Detterman, D. K. (1992). The case for the prosecution: Transfer as an epiphenomenon. In Detterman, D. K., & Sternberg, R. J. (Eds), Transfer on Trial. Norwood, NJ.: Ablex Publishing.

Dunbar, K. (1993). Concept of Discovery in a Scientific Domain. Cognitive Science, 17, 397-434.

Dunbar, K. (1997) How scientists think: On-line creativity and conceptual change in science. In Ward, T., Smith, S., & Vaid, J. (Eds), Creative thought: An investigation of conceptual structures and processes. Washington D.C.: American Psychological Association.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. Cognitive Science, 12, 1-48.

Klahr, D., Dunbar, K., & Fay, A. L. (1991). Designing experiments to test 'bad' hypotheses. In J. Shrager & P. Langley (Eds.), Computational models of discovery and theory formation (pp. 355-401). Sam Mateo, CA: Morgan-Kaufman.

Klahr, D. & Fay, A. L., Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. Cognitive Psychology, 25, 111-146.

Kuhn, D., Schauble, L., & Garcia-Mila, M., (1992). Cross-domain development of scientific reasoning. Cognition and Instruction, 9, 285-327.

Schraagen, J. M. (1993). How experts solve a novel problem in experimental design. Cognitive Science, 17, 285-309.

Schunn, C. D., & Anderson, J. R. (1999) Acquiring expertise in science: explorations of what, when, and how. In Crowley, K., Schunn, C.D., & Okada, T (Eds.), Designing for Science:Implications from Professional, Instructional, and Everyday Science. Mahwah, NJ: Erlbaum.

Singley, M. K., & Anderson, J. R. (1989). The transfer of cognitive skill. Cambridge, MA: Harvard Press.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in mental function upon the efficiency of other functions. The Psychological Review, (8), 3, 247-261.

Trickett, S., Trafton, G., Schunn, C., & Harrison, A. (2001). "That's odd!" How scientists respond to anomalous data. Cognitive Science Society Conference, 2001.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. Child Development, 51, 1-10.

Voss, J. F., Blais, J., Means, M. L., Greene, T. R., & Ahwesh, E. (1986). Informal reasoning and subject matter knowledge in the solving of economics problems by naive and novice individuals. Cognition and Instruction, 3, 269-302.

Zajchowski, R., & Martin, J. (1993). Differences in the problem solving of stronger and weaker novices in physics: knowledge, strategies, or knowledge structure? Journal of Research in Science Teaching 30, 459-470.