Human discovery of laws and concepts; An experiment *

Jan M. Żytkow
Computer Science Department
Wichita State University
Wichita, KS 67208, USA
zytkow@wsuiar.wsu.ukans.edu

Anna N. Żytkow
Institute of Astronomy, Madingley Road
Cambridge CB3 0HA, England
anz@ast-star.cam.ac.uk

Keywords: law discovery, concept generation, experimentation heuristics

Abstract

In order to understand the relationship between human and machine discovery, it is necessary to collect data about human discoveries which can be compared with machine discovery performance. Historical data are difficult to obtain, are very sparse, and arguably do not reflect a typical human performance. We contend that cognitive experiments can provide meaningful data, because the discovery tasks can be carefully defined and tailored to the comparison task, and the subject selection can be controlled in different ways.

We describe an experimental study of the human processes of concept formation and discovery of regularities. In our experiments human subjects were allowed to interact with three world models on the computer. Our results demonstrate that humans use heuristics similar to those used in computer discovery systems such as BACON or FAHRENHEIT on comparable tasks which include finding one dimensional regularities, generalizing them to more dimensions, finding the scope of a regularity, and introduction of intrinsic concepts. Virtually all our subjects made some relatively simple discoveries, while some of them were able to develop a complete theory of simple world models. The progress made by human subjects on comparably simple tasks was impeded significantly when the domain became richer, as measured by the number of regularities, their dimensionality, and the number of intrinsic concepts involved. This can be called a contextual complexity phenomenon. Our subjects demonstrated a definite pattern of chaotic experimentation and lack of theoretical progress when the level of complexity was too high.

*The work described in this paper was supported by the Office of Naval Research under grants #N00014-88-K-0226 and #N00014-90-J-1603.

The task

Little systematic study is available on human attempts at discovery tasks (Mynatt, Doherty, and Tweney, 1977; Shrager, 1985, 1987; Qin and Simon, 1990). Studies that concentrate on concept creation, such as Bruner, Goodnow, and Austin, (1956) do not consider concepts in the context of law discovery. Discovery has traditionally been perceived as an exceptional achievement that happens rarely and to special people. Recent experience with computer discovery systems demonstrates that the capability for making discoveries can be decomposed into a number of different "small" capabilities, each devoid of mystery and applicable to a broad range of cases. Relatively simple discovery systems like BACON (Langley, Simon, Bradshaw, and Zytkow, 1987) or FAHRENHEIT (Koehn and Zytkow, 1986) are successful in solving simple discovery tasks and in combining them into more complex discoveries of concept and law. FAHRENHEIT has been applied in controlling chemistry experiments and collecting data in a real laboratory, and to infer theories based on those data (Żytkow, Zhu, & Hussam, 1990). Studies on human subjects done by Shrager (1985, 1987), Qin and Simon (1990), and casual experiments reported in Langley et al. (1987, p.53) demonstrate similar discovery behavior in humans. It is reasonable to conjecture that many humans are endowed with the skills needed to become successful discoverers. World models which can be explored both by humans and discovery systems provide a controlled environment for studying human discoveries.

In order to construct and justify a framework for describing human discoveries, it is necessary to collect a lot of data in exploratory mode. In our experiments we wanted to see how humans combine different discovery skills, and how they select methods appropriate to different tasks, in particular the tasks which involve the accumulation of several discoveries. We were able to keep our subjects motivated up to several hours. Thus, simple multidiscovery tasks that could be concluded in a matter of hours (covering perhaps a few days in separate sessions) were chosen. In the future we foresee longer experiments.

Our subjects were confronted with computer simulations of simple world models. These world models were interactively encoded, allowing subjects to conduct experiments of their choice. Yet, how does one create a world of appropriate difficulty? We did it empirically, by trial and error. In the first series of experiments (section on the ZOO), we underestimated the difficulty of discovery problems. In effect, our subjects discovered very little, and the protocols were not useful for comparisons because different people explored different small subspaces of all possibilities. We have learned several interesting things, however, which we briefly report in the section on the ZOO experiments. The next two experiments (GOBLINS and GREMLINS) were more successful and we report on them in this paper.

The method

Our subjects were computer science graduate students and undergraduate seniors who had enrolled in machine learning courses. They were told that their participation in the experiment would help them understand the problems of machine learning and discovery. Although there were no time limits set on their discovery task, they were asked to continue their research as long as needed to develop a complete theory of each world, a theory that allows one to predict the result of any experiment. They were allowed to conduct as many sessions as they wished, and to resume them when they wished. Only after all subjects finished their work, the issues of machine discovery have been addressed in class. To explore the experimental worlds, the students were placed in a loop in which they designed an experiment, were provided with a result, designed the next experiment, and so forth. Each experiment took a few keystrokes, and the result was immediately available. Students developed their theories on paper.

We did not want to create abstract problems and, at the same time, we wanted to minimize the interfering background knowledge. The world models were introduced in brief stories. The stories introducing GOBLINS and GREMLINS are quoted in the next two sections. After completing their experiments, many subjects acknowledged that they quickly forgot about the stories, and started to think in terms of abstract problems. The worlds of GOBLINS and GREMLINS required at least several hours of study before a subject could develop a more or less complete theory. The level of difficulty was a little too high, but the ability to keep people interested was just about right for the selected group of subjects.

Two types of protocols were kept. First, each experiment was automatically recorded on a computer: time of experiment, input data and the result were stored. Secondly, we asked students to record as many ideas that came to their mind as possible, including all hypotheses that they entertained, and to mark the time at which they occurred. The information about time was helpful in cross-referencing both types of protocols. As we expected, the notes of students differed considerably. Still, we did not want to specify in detail what they were supposed to record. We felt that any instruction that would result in a more uniform collection of data would provide hints that might influence their discoveries.

After the students' protocols were turned in, only some clarifying questions were asked. In this way we sought explanation of unfinished phrases and obscure formalisms that they invented. No additional comments were accepted that could reflect knowledge gained at a later stage. As a result, the protocols were quite clear and adequately complemented the listings of experiments collected on the computer. We believe that the open format we selected best fits our exploratory studies. However, other methods could also be used (Ericsson and Simon, 1984).

The experiment with GOBLINS

The following story introduced the world of GOBLINS:

There are many goblins in the fairyland. Each goblin is characterized by intelligence and appetite. Both can be ANY number, and are limited to numbers. Each goblin has its energy, too, that is measured in numbers. Does energy depend on intelligence and appetite? If so, in what way? This is the problem for you. Find a complete theory, applicable to all goblins. You can develop your theory as a result of experimentation with goblins. You can also make hypotheses if you wish.

The subjects could study the world of GOBLINS by making "experiments". Each experiment consisted of selecting numerical values for intelligence i and appetite a for which the value of energy $e(i,a)$ was returned. The world model of GOBLINS is described by a numerical function $e(i,a)$ of two independent numerical variables i and a:

$$e(i,a) = \begin{cases} (a-1)(0.5i+1) & \text{for } a > 2i \\ i * a^2 & \text{for } a > 10 - i^2 \\ a^2 + 2 & \text{in all other cases} \end{cases}$$

Possible discoveries include (1) three regularities for energy $e(i,a)$ as a function of intelligence i and appetite a, that hold in three areas of the 2-dimensional plane (i,a), and (2) two boundaries that separate those areas. Each regularity can be discovered by BACON and FAHRENHEIT. FAHRENHEIT can also discover the boundaries $a = 2i$ and $a = 10 - i^2$ that separate the areas of different regularities.

Eleven students worked on this problem. The number of experiments varied between 71 and 1227. The

Table 1: Summary of experiments and theories generated within the GOBLINS experiment

Number of experiments	1227	1067	300	289	95	224	137	71	180	80	96
Formulae discovered	3	3	3	3	3	3	2	2	1	1	0*
Boundaries found	2	2	2	1	1	0	1	0	0	0	0

* Two regions of different behavior were discovered

time spent on experiments varied from 16 minutes, corresponding to 137 experiments, to 924 minutes, corresponding to 1067 experiments, plus additional time on building theories. Everybody had some success in making discoveries (cf. Table 1). Three students developed a complete and correct theory, including each of the three regularities and their boundaries. Columns 3 and 5 of Table 1 stand out in comparison to other entries because those subjects needed fewer experiments to achieve successful results. In both cases, early experiments were centered on a particular idea. It is not really possible to generalize the typical route to success, because the protocols reveal a variety of approaches. For instance, for the entry in column 6 with 224 experiments, the records show that the first conclusions were reached after one experiment, and all subsequent experiments were very systematic with a good range of positive and negative numbers as input data. The first correct formula was thought up after about 15 experiments, the second after 93 experiments and the third after 144 experiments. A long time was then spent in trying to understand the boundaries for each of the three formulae; this issue, however, was never fully resolved. Much value was attached to the predicting power of new ideas.

Protocols of students' data collection reveal a frequent use of two experimentation patterns which can be described as (1) sequential and (2) boundary search, usually bisectional. The first pattern includes all sequences of experiments in which one variable is varied whereas other variables are kept constant. A sequence of experiments can be further characterized by the initial value of the varied variable, the increment, and the number of experiments in sequence. Typically, a student kept one of the variables, say appetite, constant and varied intelligence in a sequence of 4-6 experiments, using a small integer as a starting value, and a small increment, usually 1. Then he varied appetite by a small increment and made another sequence of experiments by varying intelligence. This can be called "recursive sequential experimentation" and it occurs when the same pattern is used for varying other variables. Both BACON and FAHRENHEIT use this strategy for data collection.

Sequential experimentation was used in the search for regularities, whereas the second experimental strategy was used almost exclusively in the search for a boundary after a regularity had been detected. The bisectional search narrows the distance between two values of an independent variable (other variables held constant): the last known value for which a particular regularity holds, and the first for which that regularity fails. FAHRENHEIT uses this strategy, systematically dividing the difference in half, and terminating when the difference is less than the experimental error. Some of our subjects terminated that search when the difference was one, but some others carried the divisions to the limit of the floating point operations.

Experiments with GREMLINS

The world of GREMLINs was described in the following way:

There are many gremlins in the fairyland. There is a distinct species of gremlins for each character from a to z. For instance 'c' gremlins, or 'v' gremlins. Each gremlin belongs to one species, while there may be many gremlins in one species. Gremlins like to eat, and each gremlin has its (or his, or her) taste denoted by a positive number. Gremlins like to fight each other, too. Find the law that for ANY TWO gremlins, tells which one wins if they fight one against the other. You can develop your theory in result of experimentation with gremlins. You can also make hypotheses if you wish.

The subjects specified the type and taste for both gremlins: a letter l_i, $(i = 1, 2)$, and a positive number t_i, $(i = 1, 2)$ for each, and in return received a message about win, tie or loss by the first gremlin. Each pair (l_i, t_i) defines a particular gremlin. Invisibly to the experimenter, each letter is mapped to a number:

```
a b c d e f g  h  i  j k  l  m  n  o p
2 3 4 5 6 8 9 10 11 12 1 12 12 12 12 4.6666

 q  r  s  t u       v  w  x  y  z
12 12 12 12 8.3333 12 12 12 12 12
```

Then, the products $n_i t_i$ are computed and compared. If $n_1 t_1 > n_2 t_2$, then the first gremlin wins; if $n_1 t_1 = n_2 t_2$, then both tie; if $n_1 t_1 < n_2 t_2$, then the first gremlin loses.

Two major discoveries include (1) the introduction of an intrinsic property for each gremlin, defined by a mapping from letters to numbers, and (2) finding

that there is a boundary between the wins and loses, and that $n_1t_1 = n_2t_2$ determines the tie, while two inequalities determine the win and loss.

Fourteen students worked on this problem. The number of experiments performed varied between 55 and 589, with the amount of time spent at a terminal between 62 minutes (corresponding to 60 experiments) and 460 minutes (corresponding to 482 experiments). Three persons developed a theory that accurately predicts all experimental results, each theory based on a different formalism, in 179, 326, and 589 experiments respectively. Three other students developed almost the full theory, but were plagued by errors in assigning numbers to letters. Seven students assigned numerical values to letters or to pairs of letters. Seven never tried to map letters to numbers.

Contrary to the world of GOBLINS, where everybody detected a meaningful regularity, GREMLINS presented insurmountable problems for those who did not introduce an intrinsic numerical property. Those subjects detected only "weak" regularities, such as "the higher letter wins" (at the conclusion of 55 experiments), and "species in category 1 are stronger than species in category 2; in the same category it requires higher taste to beat someone farther away alphabetically" (at the conclusion of 125 experiments).

An intrinsic property can be introduced in various ways. Our subjects used three ways of assigning numbers to letters: (1) a number to each letter, (2) a fraction to a pair of letters, (3) express each letter by a multiplier for a selected single letter, say k: a is 2k, b is 3k, and so forth. Within (1) two mappings were used. One was the *ordinal* mapping of letters to consecutive numbers: a-1, b-2,c-3,...,z-26. The other was identical or close to the actual mapping used in GREMLINS, and described above. Below we give a sample of various experimental strategies, formalisms, and theories related to the introduction of intrinsic concepts by individual subjects.

S-1 introduced the ordinal mapping after 31 experiments, then switched to the mapping derived from experiments, but still some remnants of the original mapping remained in the final solution reached after 164 experiments.

S-2 introduced an almost correct mapping at the end of his study, after 258 experiments, as a result of theoretical analysis.

S-3 performed about 100 experiments and then introduced a "shifted" ordinal mapping, by assignments: a-2, b-3, c-4,..., z-27. Later he realized that k is special, and letters past k are equivalent, but he never gave up the initial assignment, producing as a result a very complex theory of 21 rules. The cost of maintaining a wrong paradigm is obvious, but the result was empirically adequate, with some exceptions.

S-4 performed 150 experiments and then started to use fractions of letters, and equated them with fractions of numbers, for instance $a/b = 3/2$. Very soon S-4 realized that $\frac{a}{c} = \frac{a}{b}\frac{b}{c} = \frac{3}{2}\frac{4}{3} = \frac{2}{1}$, and quickly generalized this computation to all letter fractions. Afterwards, S-4 used only fractions of consecutive letters. Finally, after 326 experiments, S-4 derived a theory that perfectly predicts results of all experiments, but never attempted a mapping to single letters.

S-5 wrote down the ordinal mapping entirely apriori, before starting any experiments. But in 178 experiments he never used that mapping and he did not produce any regularities, ending up "frustrated at the fact that the system contained so many inconsistencies."

S-6 assigned fractions after 28 experiments, then after 54 experiments switched to a "multiplier" for each number, at the same time producing a theory that perfectly predicts all results. From then, until he completed 179 experiments, he concentrated on assigning numbers to letters.

S-7 came up with another perfectly predicting theory, making 589 experiments, and using still another formalism. All letters were grouped and all groups were ordered according to different multipliers ¡of k. For instance b belongs to 2k, c to 3k. k plays a "privileged" role and does not belong to any group. No attempt to assign numbers to letters was made.

Recovering from the erroneous mapping of letters into numbers (the ordinal mapping) provided us with interesting accounts of scientific crisis and theory revision. S-3 responded to the empirical inconsistencies with complicating his theory, while S-1 changed his theory gradually. The first law had the form $t_2 = 2t_1/(N_1 + N_2)$, and was very nicely confirmed within the first 39 experiments. Then, in response to the growing counter-evidence (around experiment 93), S-2 introduced another law, not changing the ordinal mapping. $(N_1 + 1)t_1 = (N_2 + 1)t_2$. This law works nicely for letters a through e. After that, rather than creating new "small" laws as did S-3, S-2 used the new law for assigning numbers to letters. He eventually changed his law to the final form: $N_1t_1 = N_2t_2$. His protocols offer a captivating example of gradual recovery from a false theory.

Two very useful experimental strategies were introduced. One consisted in using the same taste value for both gremlins, while the letters were varied. This allowed for efficient ordering of all letters according to their "strength". S-2 used it in 84 experiments, before switching to another strategy which he used for 86 final experiments. The second strategy, used by all six successful subjects can be described as a search for the boundary between win and loss, in a sequence of experiments where two letters and one number are fixed. The second number, usually the taste of the first gremlin was varied until a draw was produced, or until the boundary was determined with the integer accuracy. The assumption that the tie is a state be-

tween win and lose was responsible for the bisectional boundary search at the beginning, while the method was later strengthened by its successful results. Several unsuccessful subjects also used this strategy, but they tried it unsystematically, abandoning it for more chaotic experimentation. Only the successful subjects systematically varied three other variables in the four-dimensional space of all experiments. Two discoveries made by all successful subjects reduced their experimentation. First, varying the other number was quickly abandoned when the law of proportionality for the tastes was detected. Then, after the symmetry between the first and the second gremlin was detected, or a mapping between letters and numbers, only selected combinations of letters were used.

The ZOO experiment

This experiment was chronologically first. The task was too difficult, however, allowing the subjects to make only sparse discoveries. The ZOO world encodes Black's law and is isomorphic to the processes in which the temperature of equilibrium is reached for a mixture of two samples (each can be water or mercury), characterized by their initial temperatures and masses. Intrinsic concepts include specific heat, latent heat, and the phase. Each experiment requires a choice of value for six variables: one nominal and two symbolic for each sample. The dimensionality of regularities and their boundaries is much higher than in the case of GOBLINS and GREMLINS. The cover story talked about animals rather than samples. Curiosity played the role of mass, while intelligence doubled for temperature. Animals stayed in pairs, and in each pair the intelligence of both animals became equal. The task was to find the intelligence of animals in a given pair.

Eight students, different from those who worked on GOBLINS and GREMLINS, worked on the ZOO problem. Nobody developed the full theory governing the ZOO (i.e. discovered the full form of Black's law) or even came close to discovering the latent heat law, the specific heat concept, etc. Some lesser things were discovered though. For instance, several students came across the numbers 0 (freezing point of water), 100 (boiling point of water), 357 (boiling point of mercury), and noted their significance, describing them as *magic numbers* (2 persons), telling that "Strange things happen for 357, 100", or that "some ranges of numbers always seem to give you the intelligence of 357". Only three persons attempted to fit some formulae to their experimental results. The rest gave qualitative descriptions only: "beta animals seem to be more intelligent", and the like.

Two phenomena differentiate the work on the ZOO from the simpler domains of GOBLINS and GREMLINS. First, it seems that the overall difficulty encountered in the ZOO impeded even partial results. It was much easier to find regularities and concepts if they occurred in the simpler worlds of GOBLINS and GREMLINS, even if they were equally difficulty as many partial results in the ZOO. The second phenomenon refers to the experimentation in the ZOO world. Everyone started with a very systematic, sequential experimentation pattern (see Section on GREMLINS), which for all but two subjects quickly degenerated into chaos. Four subjects showed a pattern of random choice of input parameters in at least the last 15% of their experiments. Chaotic experimentation patterns seem to be due to the complexity of the problem and should be further investigated as a response to non-progressing research, which does not provide new regularities and generalizations of the existing laws. Chaos emerged towards the end of most ZOO experimentation, whereas only two students working with GREMLINS started to show a bit of chaotic pattern towards the end of their efforts, complaining about complexity of the problem: "this is definitely getting complex".

Dimensions of discovery

The results of our study can be analysed along many dimensions. In the current section we will briefly summarize selected results.

Experimentation strategies: We have already reported on these in the sections on GOBLINS, GREMLINS, and ZOO. The exploration of all worlds benefited from the use of orderly experimental methods, even though some people needed hundreds of experiments to achieve success. In particular, the strategy of two persons who best described the world of GOBLINS can be characterized as that of a calculator programmed to check every detail step by step as opposed to reaching any conclusions in a more daring way.

Experimental Limitations: Most subjects used only a very limited range of input values in experimentation. Small integers ($n < 13$) were used almost exclusively. Six persons never used fractions. Most of the others used them sporadically if they could not complete the bisectional search otherwise. Five people never used negative numbers in their experiments with GOBLINS. Small integers were almost exclusively used as increments for generating experimental sequences.

Testing: Very little time was spent on testing the discovered theories. We noted an intriguing phenomenon. After making a major discovery, a student commonly switched to another task rather than going on to test the validity of the discovery. The same phenomenon can be noticed in protocols from experiments conducted by Shrager (1985). Positive thinking prevailed over critical thinking, and there is little evidence that falsified theories were rejected. In general, subjects did not change their theories in response to contradicting experiments, even if performed immediately after-

wards. Positive evidence, however, was remembered for a long time.

Complexity: First conclusions were drawn very quickly, on average after 10 experiments. This is compatible with discovery systems, which also draw conclusions that involve one independent variable after examining only a few results. Another phenomenon was quite clear: If two laws of similar complexity hold in two world-models, wm_1 and wm_2, wm_2 being overall more complex than wm_1, then it is far more difficult for human discoverers to find such a regularity in wm_2 than in wm_1. Discovery systems can succeed equally well in both models, if they are put on the right track.

Meta-principles: The modest support sufficient to draw meta-level conclusions was remarkable. Subjects who made metalevel statements used just one or few experiments to claim symmetry ("order of pairing is unimportant"), time independence ("time is not a factor"), or determinism ("no randomness in the domain"). It is actually true that both domains were deterministic, and the world of GREMLINS is symmetric.

Beyond empirical equivalence: Additional criteria can be applied to further rank the final theories. Three subjects reached the experimentally complete theory of GREMLINS, but one result ranks higher because each gremlin was represented by a number, whereas in the remaining two a number is associated with each pair of gremlins. The first theory is more comprehensive. It implies both theories of the second type, but not the other way.

Conclusion

Our research confirms that many students have been able to combine useful experimentation strategies with elementary discoveries, and to combine elementary discoveries into more complex theories. This confirms conclusions of Shrager (1985), and Qin and Simon (1990). Similarly to Qin and Simon, we found many analogies between human subjects and discovery systems in the use of operators and heuristics However, it is difficult to go beyond superficial comparisons to related work because the space of possible experiments on discovery is enormous, and each existing study explores an "orthogonal" direction. Qin and Simon (1990), for instance, used Kepler's third law to study in depth the process of one-dimensional equation finding. Although discoveries in our worlds involved several one-dimensional equations, none of them has been complex enough to allow comparison to the results of Qin and Simon. Moreover, while Qin and Simon provide students with data, we require them to make experiments. Experimentation would simplify the problem, because it would be easier to find Kepler's law in data obtained for small integers, which the subjects would undoubtedly use.

Our work can be extended in many directions. We would like to better understand the mechanisms by which humans introduce intrinsic variables and use symmetry heuristics. BACON's mechanisms for both (Langley et al., 1987) were not used, but BACON's symmetry operators require a particular experimentation strategy (Langley et al., 1987, p.171-4). If that strategy is not followed, the comparison is difficult. Experiments on new versions of GREMLINS may help in solving these problems.

Another interesting problem is the proportion of effort spent on experimentation and theories. In a yet unreported experiment, we found that even a few seconds' delay in computer response to the requests for experimental results stimulates subjects' interest in theory formation, limiting the number of experiments they perform before they start building theories.

Acknowledgements. Many thanks to Mary Edgington, Lance Petrie, and two anonymous reviewers for their helpful suggestions.

References

Bruner, J.S., Goodnow, J.J., & Austin, G.A. (1956). *A Study of Thinking.* New York, NY: Wiley.

Ericsson, K.A., & Simon, H.A. (1984). *Protocol Analysis; Verbal Reports as Data.* Cambridge, MA: MIT Press.

Koehn, B.W., & Zytkow, J.M. (1986). Experimenting and theorizing in theory formation. In: Ras Z., Zemankova M. eds.*Proceedings of The ACM Sigart International Symposium on Methodologies for Intelligent Systems* New York, NY: ACM SIGART, 296-307.

Langley, P., Simon, H. A., Bradshaw, G. L. & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes.* Cambridge, MA: MIT Press.

Mynatt, C.R., Doherty, M.E., & Tweney, R.D. (1977). Confirmation bias in a simulated research environment: an experimental study of scientific inference. *Quarterly Journal of Experimental Psychology* 29: 85-95.

Qin, Y., & Simon, H.A. (1990). Laboratory replication of Scientific Discovery Processes. *Cognitive Science* 14: 281-312.

Shrager, J. (1985). Instructionless learning: Discovery of the mental model for a complex device. Ph.D. diss., Dept. of Psychology, Carnegie-Mellon Univ. Pittsburgh, PA.

Shrager, J. (1987). Theory change via view application in instructionless learning. *Machine Learning* 2:247-276.

Żytkow, J.M., Zhu, J., & Hussam, A. (1990). Automated Discovery in a Chemistry Laboratory, *Proceedings of the AAAI-90.* AAAI Press, 281-287.