Learning Structured Systems from Imperfect Information

by

Boris Goldowsky

Submitted in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Supervised by

Elissa Newport

Departments of Psychology and Computer Science
The College
Arts and Sciences

University of Rochester
Rochester, New York

1995

Curriculum Vitae

The author was born in Southampton, New York on August 2, 1966. He attended the Massachusetts Institute of Technology from 1984 to 1988, spending the 1986–1987 school year at the University of Fribourg, Switzerland. In 1988 he received the Bachelor of Science degree in Cognitive Science from MIT.

He entered the University of Rochester Department of Psychology in 1988. As a member of the interdepartmental Cognitive Science program, he pursued research and classwork in the departments of Psychology and Computer Science. His primary line of research, conducted under the direction of Professor Elissa Newport, has been on mathematical and computer models of language acquisition.

He received a National Science Foundation predoctoral fellowship for 1989–1992. This allowed him in 1990–1991 to visit the University of Hawaii, and conduct research in comparative learning and cognition under the direction of Professor Herbert Roitblat.

At the University of Rochester, he served as a teaching assistant for courses in Psychology and Cognitive Science, and as the teacher of "Lisp Programming for Cognitive Science Students.". He received the Master of Arts degree in Psychology in 1993.

Acknowledgments

I am very grateful for the help of many people at all stages of this project, including the following:

⋆ My wife Cynthia Dill, for helping me to sort through the tangled outlines—not only of my dissertation, but of our life;

⋆ My parents, Norman Pickering and Barbara Goldowsky Pickering, and my brother Alexander Goldowsky, for never doubting that I would succeed;

⋆ Linda Dinnocenzo and the teachers and students at Brookview and Listwood Schools, for making the research possible by cheerfully sharing their space and time with me;

⋆ Elaine Kuchman, Maryann Gilbert, and Pat McLane for making everything official in record time;

⋆ Andrea Zukowski, for drawing the lilypads;

⋆ Rebecca Webb, Raphael Klorman, Mary-Jane Henning, and Jenny Saffran for helpful discussions of these ideas;

⋆ Thomas Bever, for good advice on very many occasions;

⋆ Henry Kyburg and Richard Aslin, for their participation as members of my committee;

⋆ and last but nevertheless most, my advisor, collaborator, and friend, Elissa Newport, for her indispensable help at every stage of my career as a graduate student.

Finally, although I can not list them all, I am not forgetting the many others who helped by being pilot subjects, proofreaders, cooks, consultants, housemates, classmates, and friends. Anyone who has worked on a project of this size knows that it could not have been done without the direct and indirect help of many people. Thank you all!

Abstract

A critical way in which language acquisition has been said to differ from other learning is in the results of inconsistent input. In language acquisition, it is common for the data available to the learner to be imperfect, but this has little effect on children's learning—they force the system to be regular by extracting or creating rules. In other domains, however, such as probability-learning experiments, people are said to be faithful to the probabilistic structure of the data to which they are exposed. This leads to the assumption that language must be learned by different mechanisms.

However, previous studies may have confounded the effect of linguistic *vs.* non-linguistic tasks with the complexity or structure of the learning problem, or with the age of the learners studied. Following Newport's *Less is More hypothesis,* we suggest that the structure of the system being learned interacts with the capabilities and constraints of the learner to create the different learning patterns, independent of the domain.

This dissertation describes three experiments, which use a new paradigm for presenting multidimensional structured systems. The subjects are to learn a mapping relation between objects and actions by watching observation trials which exemplify the regularities but also contain exceptions. Adults and 7-year-old children are tested on various systems which vary in complexity and quality of data, and their learning is evaluated by prediction and generalization tests.

While adults turn out to be better at learning the details of the systems, children demonstrate a tendency to respond according to a consistent pattern, inventing one if necessary. This is parallel to some phenomena of language acquisition, suggesting that language may be learned by the same mechanisms that support the learning of other complex systems.

Table of Contents

List of Tables

List of Figures

1 Introduction

PERFECTION is rare, and perfect data are seldom encountered. Most of what people are able to observe in the world around them is inconsistent, incomplete, and uncertain—and yet, these are the data that must serve as the basis for our learning, beliefs, and actions.

A prime example is the learning of a language. Children are able to acquire fluent language under widely varying conditions, even if the linguistic input that they are exposed to is sparse or inconsistent (Bickerton, 1984; Singleton & Newport, 1993). In fact, even a child growing up in a family of prolific language-users may not get enough data to deduce the structure of language from scratch, which has led to the theory that children must possess a special purpose language-learning device and a large amount of built-in linguistic knowledge (*e.g.* , Chomsky 1975). In situations where the input available to the child is particularly poor, children systematize and expand on their data, reinventing the language—or, in some situations, creating a new language. In the standard account, the language-learning module is claimed to be responsible for this creative process.

In domains other than language, however, it is generally claimed that people are quite faithful to the data that they are learning from: if exposed to an inconsistent pattern of data, rather than systematizing it, they mimic its probabilistic nature; when asked to predict the occurrence of randomly-generated events, the distribution of their predictions quickly approaches the actual probability distribution of the events (Estes, 1972).

Thus, although learning from probabilistic data has been studied in various domains, a unified understanding of the phenomena has yet to be achieved. The linguistic domain has been set apart from other domains, and learning there has been assumed to operate by different rules. However, a review of the literatures of probability learning and language acquisition, presented below, hints that there may be more in common between the two than is generally thought. The caricatures of learning theories above are simplistic versions of claims that are already oversimplified with respect to the data.

Following the literature review, an experimental study is presented which explores the possible similarities between learning in the linguistic and non-linguistic domains, concentrating on the task of learning from probabilistic information, and looking at whether children or adults will create rules when presented with unruly data. The experiments employ a new experimental paradigm in which subjects learn various complex systems that are not linguistic, but share some crucial properties with linguistic systems. In the learning of these systems, we are particularly interested in whether subjects will respond to the probabilistic nature of the data by matching its probabilities, as has been found in many so-called *probability-learning* experiments, or by forming rules and ignoring their data's exceptions, as is found in language acquisition.

If the non-linguistic systems are learned in a language-like way, then it would suggest that language learning may in fact use some of the same mechanisms as learning in other domains, and provide support for the idea that it is not the domain but the structure of the learning problem, interacting with the capabilities of the learner, that determines how learning proceeds.

1.1 Probability Learning

The data I am concerned with in this thesis are probabilistic, in the following sense: the majority of examples to which the learner is exposed follow a particular rule, but a fraction of them violate it. For example, if you are counting vehicles on the highway, most of the examples that you encounter support the following "rule":

- Highway vehicles are cars.

However, the rule is not consistently supported, since it is also true that some highway vehicles are trucks, motorcycles, bicycles, and tractors. It is not a rule in the strict sense, but rather a statement that holds true only a certain percentage of the time (higher for some roads than others). If this percentage is high enough, however, an observer may be justified in treating it as a rule, making inferences and acting based on the belief that it is true, while still keeping in mind the fact that it is not certain (Kyburg, 1990a,b).

A parallel example from the domain of language is the English past tense. For most words, the rule is:

- To form the past tense, add *-ed*.

Thus *toss* becomes *tossed*, and *row* becomes *rowed*. However, like the rule about cars on the highway, this rule is also not a complete description of the data: there are a number of common words whose past tense is made differently: *get* becomes *got*, not *getted*. The rule is useful, despite its incomplete accuracy, since it can be used to summarize much of the data and to predict the past tenses of unknown words. Furthermore, it is learned and used by children, as is demonstrated by their occasional use of forms like "getted."

The two examples above are not quite parallel, however. Children can and eventually do learn the complete system of rules and exceptions for English past tenses, but correct prediction of highway vehicles is never possible, since there is no deterministic underlying system to be learned.

1.1.1 Probability Matching

The study of learning from data that are inconsistent was brought into the laboratory as early as 1939 by means of the binary prediction paradigm (Humphreys, 1939; Grant, Hake & Hornseth, 1951). A classic form of this experiment uses an apparatus with two lamps and two corresponding buttons. On each trial, subjects are supposed to guess which lamp is going to light up, and indicate their choice by pushing the appropriate button. One of the lamps then lights, confirming or disconfirming the prediction.

If the "correct" lamp for each trial is assigned randomly, subjects will quickly come to choose each one in proportion to how often it is correct. This pattern of behavior is called *probability matching*. It is a very strong tendency, and persists through many variations on the basic experiment. For example, if the probabilities of the two lamps are changed during the experiment, the pattern of responses will also change, eventually coming back into agreement with the prevailing probabilities (Friedman & others, 1964); if these changes in probability are regular and periodic, subjects can track them with essentially no lag (Reber & Millward, 1971. For a review of many more variations of probability-learning experiments, see Estes, 1972, 1976).

1.1.2 Maximizing

The prevalence of probability matching may be surprising, if one stops to consider what the optimal strategy would be for the binary prediction task. If the subjects' goal is to make the maximum number

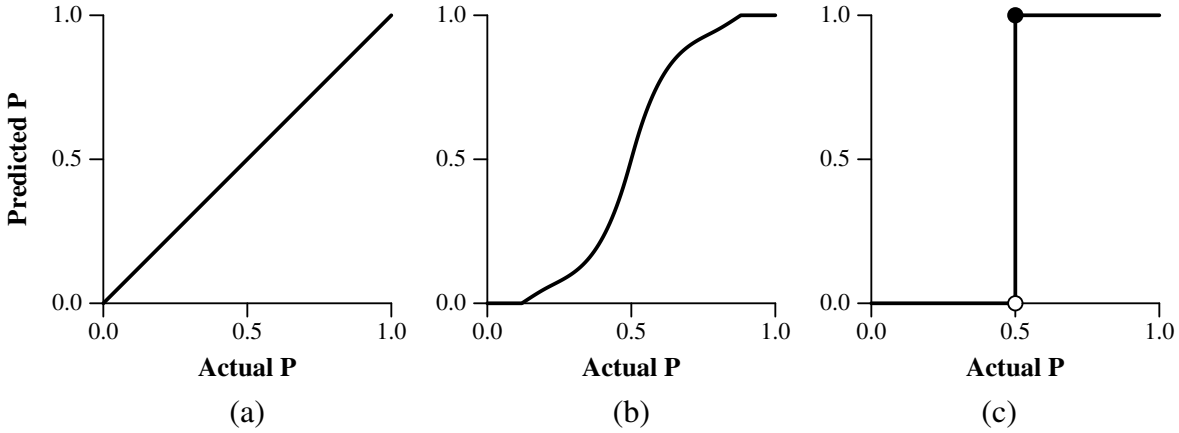

Figure 1.1: Three of the possible strategies for making predictions in a probabilistic situation. Each graph shows the probability with which the subject predicts an outcome, as a function of the actual probability of that outcome occurring: (a) probability-matching; (b) overmatching; (c) maximizing.

of correct predictions, and they know that the left button is correct 75% of the time, they can achieve the theoretically optimal score of 75% only by pressing the left button on *every* trial. This strategy is called *maximizing*. Any other strategy will result in fewer correct predictions; for example, the probability-matching strategy is correct on only 62% of trials (75% of the left-button choices plus 25% of the right-button choices).

Maximizing is rarely cited as the result of probability-learning experiments, but it does appear under certain conditions. If explicit (monetary) rewards are given for correct answers, or if incorrect answers are somehow punished, a trend towards maximizing is seen (Suppes & Atkinson, 1960; Estes, 1972). Experiments with animals, which use explicit rewards, also typically elicit maximizing. However, probability-matching may be found in animals if feedback is provided in the form of correction trials. These are "second chance" trials following an incorrect choice, in which the animal is directed to choose the correct alternative and then rewarded. Based on experiments using a correction-trial procedure, Bitterman (1986, 1991, but see Brookshire, 1978 for a critique) found that certain species will maximize (including rats, monkeys, and pigeons) and others will not (fish, and for certain cognitively demanding problems, also turtles).

Children of certain ages are reported to adopt a maximizing strategy more readily than adults (Weir, 1964; Bever, 1982): Bever's study found that 63% of 2-year-olds maximized (defined here as choosing, 6 times in a row, a cup that had candy in it on 60% of the trials), while only 17% of 4-year-olds maximized on the same task. Interestingly, another 18% of the 2-year-olds *minimized*, consistently choosing the *less* frequently rewarded cup. This minimizing response was also occasionally seen in 3-year-olds on a more difficult, 55/45% problem. Thus, although the youngest children did not always choose the correct cup, they nearly always chose a single cup and responded consistently to it.

1.1.3 Other Strategies

Probability-matching and maximizing are certainly not the only possible reactions to probabilistic data. A common intermediate strategy, called *overmatching*, is responding to the most frequent alternative with with slightly more than its true frequency. Figure 1.1 shows a schematic representation of overmatching data, in comparison to probability-matching and maximizing.

A tendency to overshoot by 3-5% has been found in a great number of two-choice probability-learning experiments with adults (Reber & Millward, 1968; Estes, 1972). One case where overmatching

is apparently the rule is in three-choice probability-learning situations; the 70% choice in a three-choice problem whose answers are correct in the proportions 70/15/15 is chosen significantly more often than the 70% choice in an otherwise identical 70/30 problem (Gardner, 1957).

Thus, it is not the case that non-deterministic data always elicit matching behavior; rather, probability-matching is a response that results, in adult humans and some other animals, from particular combinations of situation, motivation, and context.

Another example from the developmental literature may make clear the kind of subtle interactions that are at play. Weir (1964) presented subjects 3 to 18 years old with two different systems of 3-way choices. One of the response choices was correct on 33% or 66% of the trials, and the other two choices were never rewarded at all (in contrast to standard probability-learning experiments, where on every trial there is exactly one correct choice). After 80 trials, both the oldest and youngest subjects were responding, on nearly every trial, to the button that was sometimes correct. Surprisingly, however, the 11-year-olds were not. All of the subjects at intermediate ages had lower percentages of responding to the correct choice than those at the extremes of age tested.

This becomes understandable only upon analysis of the different learning styles of the three groups. The 3-year-olds were apparently responding in much the way a simple stimulus-response theory would predict: with each reward their probability of choosing the "correct" button increased, and so it quickly came to preempt the other choices completely. The 18-year-olds, however, were much slower to adopt the maximizing solution; before doing so they tried out a number of other patterns of responses—predominantly variations on left-middle-right (LMR) alternations—discarding them when they failed to work.

Weir concludes that the oldest subjects (perhaps more familiar with puzzles than with psychology experiments) came to the experiment with an expectation that there would be a "solution" which allowed them to make a correct choice on every trial, and only reluctantly gave up this hypothesis. Adult subjects' introspective reports describe the experience of being in a probability-learning experiment in much this way, as continuing attempts to out-guess the experimenter and figure out the pattern (Feldman, 1961).

The intermediate-aged children were like the adults in trying out alternation patterns—but unlike adults in that they did not subsequently reject those patterns; rather, they continued to use them despite the mounting evidence that they were incorrect (for some of Weir's groups, over half the responses were part of an LMR or RML pattern). Weir suggests that "the 7- to 10-year-old is at a point in development where his ability to generate complex hypotheses and employ complex search strategies is growing at a faster pace than his information-processing ability" (page 481), leaving them in the awkward situation of being able to hypothesize patterns that they cannot property test. Thus, they continue to blindly follow these patterns of their own creation.

Since children do not seem to follow through on the testing of the patterns they use, it is not clear whether what they are doing can truly be called hypothesis-testing (Bogartz, 1965, 1969). However, in older subjects, hypothesis-testing does become a common approach to the problem of learning from imperfect data. The hypothesis-testing subject is looking for a rule which would explain the pattern of stimuli without recourse to randomness, which is clearly preferable to one that is only right a percentage of the time. In experiments where the data actually do follow sequential patterns, adults are quickly able to find and follow them (Restle, 1967; Vitz & Todd, 1967).

Another high-level strategy which can lead to apparent probability matching is a *reward-following*, also called the *win-stay/lose-shift* strategy. Subjects using this strategy would push one button as long as it kept being correct, but whenever their guess failed, they would switch to the other possibility. This strategy is very commonly observed in animal-training situations, and is also reminiscent of one of the strategies attributed to children by Bever (1982) and others. Interestingly, the opposite strategy, which could be called win-shift/lose-stay, but is also variously named *negative recency*, *the gambler's fallacy*, or *the maturity of the chances effect*, is also found in older children (Ross & Levy, 1958). This kind of

responding reflects the assumption that random series must frequently alternate; the subject switches to a different choice any time one choice has been correct one or more times in a row.

Reward-following, in particular, is a good example of the danger of looking only at the number of responses a subject makes to each of the choices in an experiment: at that level of analysis reward-following looks exactly like probability-matching. However, it is clearly quite a different strategy cognitively, one which requires no learning at all, just a single item to be kept in short-term memory until the next trial. In general, it is difficult to deduce from a list of responses what kind of strategy a subject may be following. As an experimenter, one can only look for evidence of particular strategies, and assume that anything else is just random guessing. It is never possible to prove that a sequence is random and not actually generated by some rule that was not tested.

1.1.4 Optimality

It was mentioned before that maximizing achieves the largest number of correct predictions in the standard probability-learning experiment. However, this does not necessarily mean that it is the only logical solution to probability-learning problems—despite its name, maximizing has no exclusive claim to optimality. The notion of global optimization is difficult to define in any case, and seldom seems to be the basis of behavior (Staddon, 1980).

However, one can describe some simple situations in which several of the strategies discussed above—probability-matching, maximization, and hypothesis-testing—can be shown to be optimal.

In the typical probability learning experiment, there is exactly one correct choice on each trial, which is determined randomly and independently of any other trial or response, and which is shown to the subject on each trial by means of one of the lights going on. In this simple case, the optimum strategy is indeed maximizing—allocating all your choices to alternative that has proven correct on the largest number of past trials.[1]

On the other hand, if there are sequential patterns in the data, then a maximizing solution is clearly not optimal; what is required is some strategy that will find those patterns and follow them. The problem domain here is too vague to permit any claims of optimality, but some sort of hypothesis-testing pattern seems to be called for. What people actually do does not seem to be anything like an optimal search strategy (Weir, 1964), but it is effective.

Finally, in the absence of patterns, there are situations where probability-matching turns out to be the optimal strategy. Optimal foraging behavior has been extensively studied (see Gallistel, 1990 for a review), and offers some examples: consider a situation where there are various possible locations that food may be found, but there is a limited amount of food at each location. In this situation, it is desirable for each forager to distribute its feeding to the various locations in proportion to the amount available at each; this distribution of the load, called the *ideal free distribution*, minimizes the likelihood that any of the locations will be depleted. In addition to serving the common good, it also leaves the greatest amount of food available to each individual if the group of animals is sufficiently numerous.

An nice example of the ideal free distribution in a relatively natural setting is described by Harper (1982): two experimenters threw pieces of bread at different rates into a pond where a group of ducks were living. The ratio of the numbers of ducks waiting in front of each experimenter soon came to accurately reflect the ratio of the rates at which they were providing food. Clearly it would have been foolish of the ducks to "maximize" by all gathering in front of the experimenter who was more forthcoming.

In the laboratory, an analogous situation can be set up with a *Fixed-Interval (FI)* or *Variable-Interval (VI) Schedule* of reinforcement (*e.g.* , Herrnstein and Vaughan, 1980). These are conditioning paradigms

[1] But in a slightly different experiment like Weir's (1964), in which you can only determine if a choice would have paid off by choosing it, things get more complex because some number of trials must be spent exploring the choices in an effort to determine which one is most profitable; the optimality problem becomes one of deciding when to stop exploring and begin exploiting the favorite choice. An approach to this far more complex optimality problem is described by Gittins (1989).

where an interval of time must elapse between reinforcements of a particular behavior; in the FI schedule this interval is constant, while in a VI schedule it varies randomly around some average. If two behaviors are being taught by using independent interval schedules, then the subject must choose at each moment which of the behaviors to perform. If both schedules allow reinforcements equally often, then it is reasonable to simply alternate. If, on the other hand, the average times between reinforcements are different, say one minute and ten minutes, then a better strategy is to perform the more frequently rewarded behavior more often. However, giving up entirely on the second behavior is also a poor idea, since that would result in missing its scheduled reinforcements.

Animals put in this situation are found to obey the *Matching Law* (Herrnstein, 1970): the frequency of responses are in proportion to the frequency of reinforcements, so in the situation described above, the animal would perform the first behavior 10 times as often as the second behavior (Schwartz & Reisberg, 1991). Optimality in this situation is difficult to define precisely, but this strategy is at least an excellent approximation (Staddon, 1980).

Given the similarity (from the subject's point of view) of this paradigm to the probability-learning paradigm, perhaps it is not surprising that the same kind of responding appears in both situations. Faced with choosing one of two occasionally-reinforced responses, a subject does not necessarily know whether the experimenter is judging answers according to a table of random numbers, or according to a variable-interval stopwatch. The probability-matching strategy may be a case of a behavior that is optimal in one circumstance (*e.g.* , foraging) being used in a situation where it is not the best choice.

1.1.5 Models of Probability Learning

Herrnstein's *Matching Law* can perhaps be called the first model of probability learning. Its basic form, for two choices A and B, is

$$\frac{R_A}{R_A + R_B} = \frac{r_A}{r_A + r_B} \tag{1.1}$$

where R_A and R_B are the number of responses to choices A and B, and r_A and r_B are the number of reinforcements following such choices (Herrnstein, 1970).

The Matching Law is actually considerably more flexible than its simplicity might suggest. For example, it accounts for both the probability-matching and maximizing that is found in many animal experiments. Recall that maximizing is typically found when correction trials are not used: after each choice, the animal is rewarded if its choice was correct, or must wait for the next trial if it was incorrect. Assume that the animal initially responds equally to the two alternatives, and that A is reinforced 75% of the time. The animal would be rewarded on 75% of its choices of A, and 25% of its choices of B, so the Matching Law predicts that it would begin to choose A on 75% of the trials. The increase in responses to A, and concomitant decrease in responses to B, has the effect that an even larger proportion of the reinforcements earned by the animal now follow responses to A: now A is being chosen 3 times more often than B, and it is, as before, reinforced 3 times as often when it is chosen, so without changing the setup of the experiment, the subject is now receiving 9 times as many reinforcements after choices of A than reinforcements after choices of B. This in turn causes a further increase in choices of A, and the animal quickly approaches complete maximizing.

In a correction-trial procedure, however, any incorrect choice is followed by a forced, and rewarded, choice of the other stimulus. Every trial ends with a reward, and the proportion of those rewards is fixed at 75% A to 25% B. The Matching Law predicts that in this situation the subject's responses will likewise be fixed at 75/25.

As we have seen, these predictions are not correct in all situations or for all subjects; many factors besides feedback affect whether a subject will maximize, probability-match, or follow some other pattern. The Matching Law does well for its simplicity, but has a number of drawbacks: its dependence on explicit rewards, when in fact observation of probabilities is sufficient for learning (Gallistel, 1990;

Reber & Millward, 1968); the lack of any way to account for different contexts in which events happen; and unrealistic predictions about the early stages of learning (it predicts wild swings in the probabilities of early responses, since when relatively few rewards have been experienced, the ratio changes sharply with each one. Such swings are not observed, however; see Gallistel 1990).

The linear model proposed by Rescorla and Wagner (1972) is able to correct some of these problems, due to the fact that it separates learning about the choices into separate *response strengths*, rather than collapsing all the information into a single ratio. A reinforcement increases the response strength of the most recent choice, and the ratio between the strengths of the available choices determines the likelihood of picking them on subsequent trials. The insertion of the intermediate representation of a response strength between the subject's input and output allows Rescorla and Wagner to model more more closely the actual shape of learning curves, as well as to account for the effects of *non*-reinforcement following a choice. It is also possible to begin to look at situations where the choices are made in a context—ie, in which multiple stimuli are present, each of which may influence the distribution of rewards. With suitable assumptions about how the strengths of different items can be combined, this becomes a powerful model of choice behavior (see, for example, Bitterman, 1986, Couvillon and Bitterman, 1991).

Adding further complexity and flexibility (and also, for the first time, non-determinism in learning), William Estes' *Stimulus Sampling Theory (SST)* takes an environment containing multiple elements as the basic situation. Estes and his followers have constructed statistical models to account for a wide variety of probability-learning situations, including some very elaborate variations (Estes, 1972; Estes, 1976). There are various versions of SST, but the basic premise is that each stimulus can be considered to be made up of a set of elements. On each trial, a subset ("sample") of these elements is noticed by the subject, and their representations become active. Each of the elements can also be associated with a particular action (*e.g.* , choosing the left button, or moving to a particular feeding location), and the subject's response on each trial is determined by the connections from the currently active stimulus elements. If that response is rewarded, the active elements become associated with that response, if they were not already.

One interesting feature of SST is that it can put the randomness of responding into the perceptual system, not the act of choice itself: the response is wholly determined by the connections of the active stimulus elements, but the choice of which stimulus elements are noticed on a particular trial would seem to be a perceptual question. Combined with a mechanism of selective attention, this can be an extremely powerful model (Bower & Hilgard, 1981).

Stimulus-Sampling Theory is quite flexible, and has been applied to a wide array of probability-learning situations, but it still does not provide a single model that is able to account for all of the experimental data that has been generated. One particular failure that SST shares with all of the other models I have discussed is the lack of any account of the use of the patterned responding or hypothesis-testing (see section 1.1.3), or even an account of its own boundary conditions, which would predict when a subject would give up on statistical responding and begin to try out patterns.

1.1.6 Summary

I have outlined some evidence which shows that probability-learning is not a simple phenomenon. Rather, there is a complex array of factors at play that collectively determine how a subject will respond when presented with an array of inconsistent data. The possible styles of learning and responding include probability matching, maximizing, and compromises between those, and various forms of strategic guessing, pattern following, and hypothesis testing.

The factors that may affect the choice between these include characteristics of subjects (their species, age, and understanding of the situation), and of the task (the use of correction trials, reward or other motivation, instructions, and the cognitive demand of the problem). Furthermore, these factors interact with each other in complicated ways.

Clearly, people have many different learning styles and strategies in their repertoire, and probability-matching, far from being a reflex response to probabilistic data, is just one of a number of learning tools that can be deployed when needed. The differences between adults and children can therefore be considered as differences in the array of tools available, and the sophistication with which the proper one for the situation can be chosen.

1.2 Language Learning

Learning from inconsistent data is also a recurrent theme in studies of language acquisition, although it goes by different names and is not generally connected to the probability-learning studies discussed in the previous section. However, many investigators have looked at how learners can form rules on the basis of information that is essentially probabilistic in character. In the following descriptions of some of these areas of research, each is approached from the viewpoint of probability learning, in an attempt to construct a unified picture of the phenomena.

1.2.1 Overregularization

An important example in language acquisition, which has already been touched upon in the beginning of this chapter, is *overregularization*: the child's common error of using a regular rule too often, for example using *goed* for the past tense of *go*, rather than the correct irregular form *went* (see Marcus et al., 1992, for an extensive review).

A common interpretation of this finding is that children are willing to innovate linguistic forms, or change the forms that they have heard, in order to bring them into line with their innate predisposition to find grammatical rules. However, in this case there is an alternate interpretation, due to Singleton and Newport (1993), which looks at the problem facing the child as a probability-learning situation: often, say 75% of the time, an event that is in the past will be spoken of with an *-ed* ending on the verb; the remaining 25% of the time, past-tense verbs have a different ending or no ending. On this view, the child's behavior of over-using the majority *-ed* ending is an example of maximizing—or perhaps more appropriately overmatching, since children do not generally make overregularization errors 100% of the time.

I do not want to make too much of this example, since the analogy between trials of a probability-learning experiment and exposures to the past tense morpheme is a weak one. The exposure to past tenses is not random; there are rules that predict the occurrence of even the irregulars. Some of these rules may only cover a single verb, such as the idiosyncratic *go-went* and *is-was*, but they are still perfectly consistent in the sense that there is exactly one past tense for each verb, and eventually, the child will master all of these exception-rules. It is tempting to propose a connection between overmatching and overgeneralizing, but before committing to this idea it would be wise to look for some cases in language acquisition where the learner's input is not just varied in an orderly way, but truly inconsistent.

1.2.2 Creolization

One case where a child's language input can contain true inconsistency is when the adults in the child's environment are not native speakers of the language that the child is learning—as when a new language is first being formed out of a contact vernacular, or pidgin language.

Pidgins are created by necessity in areas of contact between groups of adults who share no common language. They are characterized by variability and inconsistency, with vocabulary items and grammatical rules being borrowed from the various native languages of the speakers (Kay & Sankoff, 1974). If this pidgin becomes the common language of the community, however, and children grow up with it as their

native tongue, it is transformed in dramatic ways. Sankoff (1979) lists seven primary changes that have occurred in Tok Pisin, a language of Papua New Guinea that has been undergoing such a transformation quite recently. Some of the changes are elaborations (e.g., marking singulars and plurals, which had not been consistently done before), while others are simplifications or streamlining (*e.g.* , a future marking adverb, *baimbai*, is shortened to */bə/* and moves to attach to the verb: Sankoff and Laberge, 1973).

Derek Bickerton has claimed that Tok Pisin shares with all other creole languages a universal core grammar, which is essentially the minimum requirements that children demand of their language, and which children are willing and able to create if it does not already exist. His *Language Bioprogram Hypothesis* further claims that "the most cogent explanation of this similarity is that it derives from the structure of a species-specific program for language, genetically coded and expressed, in ways still largely mysterious, in the structure and modes of operation of the human brain" (Bickerton, 1984, page 173).

On the other hand, Singleton and Newport suggest that the morphological and syntactic rules added to the language may not have been made up by the children from nothing; rather, they may have been present in the adults' speech, but used inconsistently (Singleton and Newport, 1993; also see Sankoff and Laberge, 1973, for relevant data). As in the case of *baimbai*, children may be noticing common but inconsistent patterns, and transforming them into obligatory markers. Regardless of whether this hypothesis can account for *all* of the changes that go on in creolization, it is undoubtedly the case that many changes are of this type: a word, morpheme, or structure used somewhat frequently becomes the obligatory element, to the exclusion of any other method of marking that meaning which might have been used in the community.

This probabilistic character of pidgins, then, is another possible analogy to the probability-learning paradigm. Just as in overregularization, we have a case where children's behavior can be described as selectively ignoring parts of their input in favor of a bias towards systematicity (Newport, 1981, 1982; Singleton and Newport, 1993). In overregularization, this is restricted to the occasional use of a regular rule in a novel (*i.e.* , incorrect) context, and decreases with time as the irregular forms are more solidly learned, but in the case of creolization the process of selecting and systematizing is able to continue all the way until the biases become rules, and the system becomes a new language.

1.2.3 Simon

It may be a phenomenon like creolization, but on a smaller scale, which underlies Singleton and Newport's (1993) findings in their study of Simon. Simon is a deaf child who learned American Sign Language from his parents, who are non-native speakers (both learned ASL at age 15). The parents use ASL in a manner typical of late learners: their grammar and morphology are quite reduced and inconsistent. On a test of verbal morphology, they scored an average of only 65-70% correct. However, despite the fact that Simon had no other ASL input from which to learn, on the same test he scored significantly higher than his parents, and on 5 of the 7 morphemes tested his performance was indistinguishable from the performance of children of native speakers, 90 to 95% correct.

How was it possible for Simon to surpass his language model in this way? It turns out that for most of the morphemes tested, Simon's parents produced the correct form in the majority of cases, as does Simon. However, Simon very seldom produces any of the numerous errors that his parents do, so that their 65% correct responses become 90% correct at his hands. In the one case where the parents' most common form was actually an incorrect handshape for the meaning, Simon also uses this (erroneous) form consistently.

1.3 Comparisons

Phenomena like overregularization, creolization, and the learning of language from the inconsistent usages of late learners have led many researchers to the conclusion that children must have some innate predispositions for learning language, a "Language Bioprogram" or "Language Acquisition Device." There is, however, another possible explanation: that children find patterns in their input, and by regularizing those patterns elevate them to the status of rules (Newport, 1981; Newport, 1982; Singleton & Newport, 1993). Children seem biased towards systematicity, but as Newport and Singleton suggest, this is quite analogous to the phenomenon of maximizing in non-linguistic learning, which is by no means restricted to language. This is the perspective that I have adopted in the preceding discussion. Overgeneralization, creolization, and language learning from impoverished input may be, at least in part, manifestations of the child's propensity to maximize when faced with certain kinds of inconsistent data.

The intention of this study is to test this hypothesis experimentally. If the learning phenomena under discussion are caused by the capabilities and biases of the young learner interacting with the structure of linguistic systems, and not by the domain of language *per se*, then similar phenomena ought to appear when a child is given a non-linguistic system to learn that shares some crucial aspects of the structure of languages.

It is, unfortunately, not clear what aspects of the structure of languages are the important ones to include in such a system, and in the duration of an experimental session it is obviously not possible to test very many or very complex systems. Nevertheless, it is possible to take some guesses about what kinds of structures might be important, and at least to try a few different levels of complexity, ranging up to a moderately complex system. A first attempt at doing this is described in the following chapter.

Since the learning problems presented to subjects in these experiments are complex, there is a correspondingly wide range of possible responses, and there are several dimensions along which results can be measured. The most straightforward of these is the accuracy with which the subjects learn to predict elements of the system, and this accuracy data will be presented along with the procedures in chapter 2. However, as will become clear, it is also interesting to look at how internally consistent subjects are in their use any sort of response pattern, regardless of whether that pattern is accurate or not (recall the examples of innovation in language learning, and the minimizing behavior described in section 1.1.2). This consistency data, presented in chapter 3, gives quite a different impression of what the children are doing than does the accuracy data alone.

The results of these experiments suggest that children may indeed have some unique ways of reacting to inconsistency in structured systems, even in a non-linguistic domain. Although this study is much more exploratory than explanatory, some possible implications of the results are presented in the final chapter, along with ideas for further research. The paradigm introduced by these experiments proved to be quite a fruitful one, although not in the ways originally anticipated. Perhaps future inquiries along the lines sketched here will shed some light on the question of exactly what causes rule-making behavior in children's learning, both in and out of the linguistic domain.

2 Experimental Methods and Initial Results

The experiments described here are designed to look for rule-like behavior in response to imperfect information in a non-linguistic domain. The first requirement, then, is to construct a system that is not linguistic and is simple enough to be taught to subjects in a brief laboratory session, but, at the same time, which is capable of including some language-like structure.

The system that used here to accomplish this is a multi-dimensional mapping between *objects* and *actions*. The objects are colored geometric shapes, with color, shape, and size being independent dimensions of variation. The actions are also organized in terms of separable dimensions: the direction, type of path, and manner of motion. An example of a complete stimulus, then, would be a large green square, rolling on a curved path to the right.

Two important aspects of linguistic systems are represented in miniature by this system: reference and compositionality. *Reference*, in linguistics, refers to the fact that linguistic systems are connected to the world: words can refer to concepts or physical objects. The concepts and objects, meanwhile, each have their own scheme of organization and interconnections, so that the task of language acquisition includes learning connections between (at least) two complex and differently-organized systems.

In these experiments, there is also a sort of reference, but in a very impoverished sense. There are two domains taking part in the learning problem—objects and actions—and each has its own organization. In the learning of morphology, the task is to connect morphemes to their meanings; in these experiments, subjects learn to connect features of objects with features of actions. Thus, the form of the problem is the same, although it is presented here in a miniature version. The object and action domains in these experiments are extremely simple compared to the domains of morphology or semantics, but, as the data will show, they are not trivial to learn within the constraints of the experimental session.

The second language-like aspect of the experimental task is *compositionality*. Linguistic units (*e.g.* , words) are made up of smaller meaningful parts (morphemes), which combine in rule-governed ways to create complex meanings. In these experiments, it is possible to break a complete stimulus down into smaller parts (just its color, for instance) and find a "meaning" for each part (*e.g.* , the color of the stimulus might predict its direction of motion).

The learning problems in these experiments are presented as sequences of examples (with each example comprising an object and its action), but the system to be learned is actually structured like a simple morphology: if "the small blue star" is a like a word, then the color "blue" is like a morpheme of that word. If all the blue objects bounce, then we can continue the analogy by saying that "bouncing" is the meaning of the "blue" morpheme. And if this meaning combines with the meanings of the size and shape to create the fully-specified action, then we can call it a compositional system.

In contrast, a system could also be constructed in which the particular objects had characteristic actions, but there was no way to break them down into morpheme-like components—small blue stars might always bounce to the left, but without there being any consistency in the way all blue objects move. Such a system could be learned only by memorizing the behavior of each individual stimulus

item, whereas the compositional systems used here can be learned not only by memorizing individual stimuli, but also by analyzing the stimuli into their components, and finding the low-level regularities.

2.0.1 Independent Variables

Quality of Data The pivotal independent variable in these experiments is the *data quality*—whether the sequence of examples provided to the subject perfectly reflects the underlying system, or has errors introduced into it. We saw in section 1.2 that young language-learners are able to find (or innovate) rules even if their input does not follow those rules consistently. Changing the consistency of the data provided to subjects allows us to test whether the learning of an abstract system like this is affected (or unaffected) in a similar way by inconsistency. In each experiment discussed below, the same system will be given to several groups of subjects with differing levels of inconsistency in the data, ranging all the way from data that perfectly follow the rules to random data that follow no rules at all. At each level, we will look first at the subjects' accuracy (how often did they follow the pattern that the data suggested?) and later also at their consistency (how often did they follow any patterns at all?). This will give us an idea of how subjects are influenced by the consistency of the data they get about a system.

Complexity of Systems Might complexity affect the subject's reaction to inconsistent data? It is certainly plausible that a young learner, having particularly limited resources (such as memory and attention) available, when confronted by a complex system to learn, may need to simplify the situation in any way possible. Discarding all but the most frequent types of items in a particular category may be an effective kind of simplification. This is a form of Elissa Newport's *Less is More hypothesis* (see Newport, 1984, 1988, 1990, 1991, and Goldowsky and Newport, 1993): The limitations of the child learner may be responsible for their strong ability to acquire languages.

If it is sheer complexity that is responsible for children's maximizing behavior in learning languages, though, this experiment may well be insufficiently language-like to show any maximizing at all. Languages are much richer and more complex than any system that can be taught within the time constraints of an experimental session. However, the complexity of the system used can be manipulated within those constraints in order to study this question. By varying the degree of complexity of the presented system within reasonable bounds, the direction of the effect of complexity can be determined. If subjects show a greater tendency to maximize when exposed to a more complex system, or if children maximize more than adults when presented with the same system, it would provide a measure of support for the speculation that the even greater complexity of languages may be what causes the maximizing outcome.

Differences in complexity may also provide a way to unify findings about maximizing in various different contexts. In Bever's (1982) experiments, maximizing was found to be dramatically less common for subjects above 2 years of age, but overregularization of morphological rules goes on nearly undiminished to at least age 5 (Marcus, Pinker, Ullman, Hollander, Rosen & Xu, 1992). One could speculate that the age difference is due to the far greater complexity of the language problem, which remains challenging long after the child has reached adult-like performance on the two-choice probability-learning task. Indeed, Weir found that the age that shows the most maximizing is dependent on the particular problem in question (Weir, 1964)

For these reasons, the complexity of the presented systems is manipulated in this study. The first experiment uses a reasonably complex system, in which the objects and actions each are composed of 3 independently-varying dimensions having 3 possible values ($3 \times 3 \times 3$). The second and third experiments each compare two simpler systems, a two-dimensional system (3×3) and a single-dimensional system (of 3 stimuli). If complexity affects learning as Newport's and Bever's work suggest, then we would expect to see more maximizing in the more complex systems.

Age of Learners The final major variable to be examined is the age of the learners. It is known that children approach the task of learning a language very differently than adults do (see section 1.2). This implies that if the same mechanisms that are used in language acquisition are being recruited to do the present task, we should find parallel differences in adults' and children's results. For this reason, Experiments 2 and 3 compare adults and 6 to 8 year old children given the same problem under very similar experimental conditions. Again, the comparison to language acquisition would lead us to suspect more maximizing in children than adults.

2.0.2 Dependent Variables

Two major dependent variables are used to describe the data. In this chapter only *accuracy* will be considered; this is the percentage of trials on which subjects pick responses that follow the underlying system that they were exposed to—although, since the data that the subjects get to see is probabilistic, it may not have followed that underling system with perfect consistency.

In the next chapter, however, a different metric will be proposed: *consistency*. This is a measure of how well the subject followed *any* pattern in their responses, independent of whether that pattern reflected the patterns in the data or not. The consistency data, and in particular comparisons of the two measures, paint quite a different picture than the accuracy data alone. However, further discussion of this must be postponed until after the description of the experimental procedures and initial data.

2.1 Experiment 1

In this first experiment, a methodology is introduced for teaching complex systems of mappings to subjects, with varying amounts of inconsistency in the data to which the subjects are exposed. The mappings are between a set of objects and the actions that those objects perform; the subject's task is to be able to predict the action given the object. An added layer of complexity is that although the stimuli are presented individually, there are underlying regularities having to do with the dimensions of variation of the stimuli, with each object-dimension (*e.g.* , shape) mapping to an action-dimension (*e.g.* , direction of motion).

The basic situation is one in which the stimuli are constructed according to a perfect mapping between objects and actions. In the condition of greatest experimental interest, however, the mappings are similar but probabilistic; so that instead of red objects always bouncing, 76% of them bounce, and the remaining 24% are distributed evenly across the other possibilities. As a further baseline condition, for some subjects the objects and actions are paired randomly, so that knowing the object provides no information for predicting the action. This *random* condition is intended to uncover biases that subjects have toward particular patterns of responding, independent of the data provided about the system.

It is also possible to test several different levels of consistency within a single set of stimuli, because exceptions can be introduced into each dimension independently. The final condition, then, is *Mixed*. To illustrate, in one of the Mixed systems used, shapes mapped onto directions of motion perfectly, sizes mapped onto paths with some exceptions, and colors and manners of motion varied randomly and independently. This allows within-subject comparisons of the effect of consistency, as well as allowing a test of whether subjects treat the different dimensions as separate problems, or if the interactions between dimensions that are mixed in this way affect the manner in which each of the component learning problems is approached.

Approximately 1/4 of the possible stimuli in this and following experiments are not used during the training phase of the experiment, but saved for a generalization test: for example, the subject may never be shown a small red square until the last block of the experiment. However, if the subject has learned rules about the behavior of small objects, red objects, and square objects, they should be able to apply

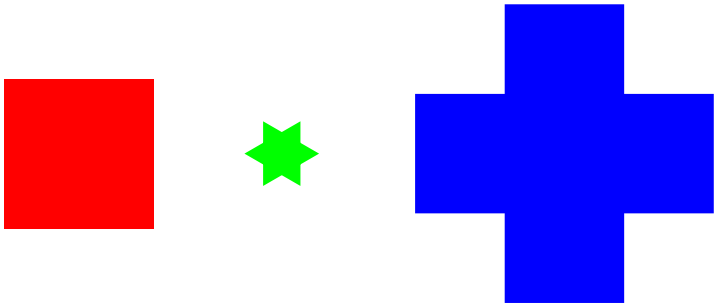

Figure 2.1: Three of the objects used for stimuli. They ranged from 1 to 4cm across, as illustrated, and came in the three shapes and three colors shown. All 27 possible combinations of these features were used in Experiment 1.

these to the case of the small red square, and correctly guess its type of motion. This result would indicate that the subject had abstracted the rules relating dimensions of the stimuli; whereas an inability to do this generalization (while still performing well on the trained stimuli) would suggest the subject had memorized individual stimuli without extracting the dimensional rules.

2.1.1 Subjects

The subjects for this experiment were 36 (22 male, 14 female) University of Rochester undergraduates enrolled in an introductory Psychology class, participating as part of a course requirement. Ages ranged from 18 to 21 years. They were tested individually by the experimenter.

2.1.2 Stimuli

All of the stimulus presentation and response collection was implemented using an Apple Macintosh Si microcomputer, running a special-purpose program written in the C programming language.

The stimuli were colored objects moving in various ways on the white background of the computer's screen. The objects varied in three of their attributes (or *dimensions*): color, shape, and size. Each dimension has three possible values: color was red, green, or blue; shape was square, star, or plus; and the size was 1, 2, or 4cm across. There are 27 possible objects that can be created by combining these dimensions; figure 2.1 illustrates a few examples.

On each trial, one of these objects was presented, engaged in one of various actions. Each action was a way of moving from the center of the screen to the periphery. The actions, like the objects, differed in three dimensions: manner of motion, direction of motion, and shape of the path. Each of the dimensions had three values: the manner of motion was rolling, bouncing, or stop-and-go (moving at double the average speed for 333ms, then stopping for 333ms); the (initial) direction of motion was 90°, 200°, or 290° (measuring clockwise from straight up); and the path was linear, curved, or angled (straight for 3/4 of its travel, then taking a 90° bend to the right). There are thus also 27 possible actions; see figure 2.2 for examples.

Although one could construct 729 (27×27) different stimuli by combining the objects with the actions, in these experiments the stimuli were constructed more selectively, with a 1:1 mapping of objects to actions being the target system that subjects were supposed to learn. Each of the 27 objects, in other words, was paired with a single, specific action (although different for different subjects), and the pairing was the information subjects were tested on. The systems were constructed as follows:

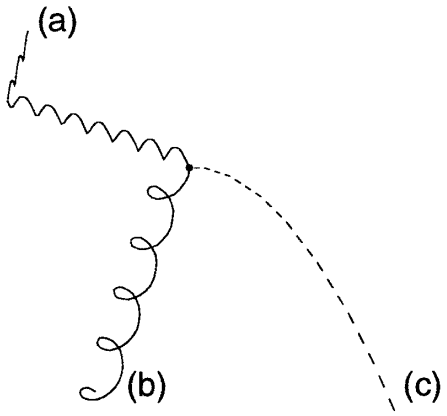

Figure 2.2: Three of the actions used for stimuli. (a) Diagonal, angled, bouncing; (b) downward, linear, rolling; (c) rightward, curved, stop-and-go motion. Each path begins at the central point and starts its motion in one of the three standard directions; the curve and angle eventually deviate from their initial direction. The schematic representations of the actions shown here are the same ones that were manipulated by Experiment 1 subjects to indicate their responses.

- Each object dimension was connected with an action dimension. This was done in three different ways in order to counterbalance any effects of the differing physical properties of the dimensions. Thus for one-third of the subjects, the size of a stimulus was related to its direction of motion; for another third, to its path; and for the rest, to its manner of motion.
- Each *value* on each object dimension was connected to a value on its corresponding action dimension. This, also, was done in three different ways as a counterbalancing measure: for example, for one-third of the subjects for whom size was related to direction, small corresponded to left, for another third medium corresponded to left, and for the rest, large corresponded to left.

The three different ways of matching up the dimensions and the three different ways of matching up the values on those dimensions define 9 different variants of the mapping system. One of these variants was given to each of the 9 subjects in each group. However, all 9 subjects in each group experienced the same *type* of mapping, as described below.

Of the 27 object/action stimuli in the system, 6 were reserved for the generalization test and were only presented in the last phase of the experiment. The remaining 21 stimuli were presented once in each block, with their order randomized independently for each block and subject.

Subjects in the *Perfect* group were exposed to this set of stimuli, exactly as described, for the entire experiment. For the Perfect-condition subjects, then, the stimuli illustrated a 100% consistent mapping between each of the object dimensions and its corresponding action dimension.

For subjects in the *Imperfect* group, however, some of the stimuli were modified. For each of the three action dimensions, five stimuli were chosen from each block of 21, and that dimension of the action of each of those stimuli was changed to one of the two values that did *not* follow the pattern. Thus 24% (5/21) of the stimuli had a mapping inconsistent with the pattern on each dimension, while 76% (16/21) were left with the normal mapping. These random mutations were done independently on the three dimensions, which resulted in an average total of 44% of the stimuli having all three of their dimensions following the pattern as originally described; 42% of the stimuli having one dimension different from that which the pattern would predict, 13% differing in two dimensions, and the remaining 1% having all

their dimensions different from their original values. The mutations were produced in such a way that each particular stimulus object had its action altered just as often as each of the others: no stimulus was more or less reliable than any other.

For subjects in the *Random* group, there was no correlation between the objects and the actions that they were shown. This disorder was created by an exaggeration of the means described above: for each dimension and block, 14 stimuli were mutated. Thus, if squares were originally associated with rightward motion, 7 of the 21 squares retained this rightward motion, 7 were altered to move left, and 7 to move down. This resulted in the stimuli giving no information about object-to-action mappings.

The fourth group, *Mixed* subjects, were given one object dimension that was perfectly correlated with an action dimension, one that was imperfectly correlated, and one that was random. The idea of this condition was to see whether learning would proceed on each dimension the way it did in the non-Mixed conditions, or, alternatively, if they would interfere or interact with each other. If there were no interaction, it would permit within-subjects examination of the effect of data quality.

2.1.3 Procedure

The subjects were randomly divided into the four groups (Perfect, Imperfect, Random, and Mixed) with the stimuli presented to each group differing as described above. Each subject was also assigned to one of the nine different systems that counterbalanced the particular dimensions and values that mapped onto each other. The instructions and all other procedures were the same for all subjects.

The experiment was divided into four sections: observation, pretest, feedback trials, and post-test. In the *observation* section, subjects were shown a sequence of examples of the objects performing actions. The instructions were to "watch carefully, and remember as much as you can about the objects you see, and the way that each of them moves." Observation trials have been found by Reber and Millward (1968) to be useful for giving probability-learning subjects a lot of information quickly, and pilot tests had shown them to be useful in this task as well. There were 6 blocks (126 trials) of observation. Each block took just over 3 minutes, and subjects were allowed to pause for as long as they liked after each block.

In the *pretest* section, the subject was shown each object (presented motionlessly in the center of the screen) and asked to guess how that object would move, based on what they saw in the observation section. Responses were made by means of a 10-key keyboard, which had a button for each of the possible directions, paths, and manners of motion, and an "OK" key. As they made their choice, it was schematically represented on the screen with diagrams like those shown in figure 2.2. They were free to self-correct until they pressed OK, at which point the screen was cleared and the next trial would begin. No feedback on the accuracy of the guess was provided. There were 2 blocks (42 trials) in the pretest.

Following this there were *feedback trials*, which proceeded identically to the pretest trials until the OK button was pressed. At this point, depending on their response, either "Yes!" or "Close" would be presented in the center of the screen for 333ms, or the entire screen would be flashed black for the same period of time to designate an incorrect response. A "Close" response was defined as having some but not all of the dimensions correct. Following this feedback, the correct answer was shown, in the same manner as it had been shown during the observation trials (except that the representation of a close or correct guess would remain on the screen for comparison with the actual action). There were 4 blocks of feedback trials (84 trials).

Finally, the *post-test* was identical to the pretest, except that it included the generalization stimuli, which had never been presented or tested before. Two blocks (54 trials) were in the post-test. Most subjects completed the entire experiment in between 60 and 75 minutes.

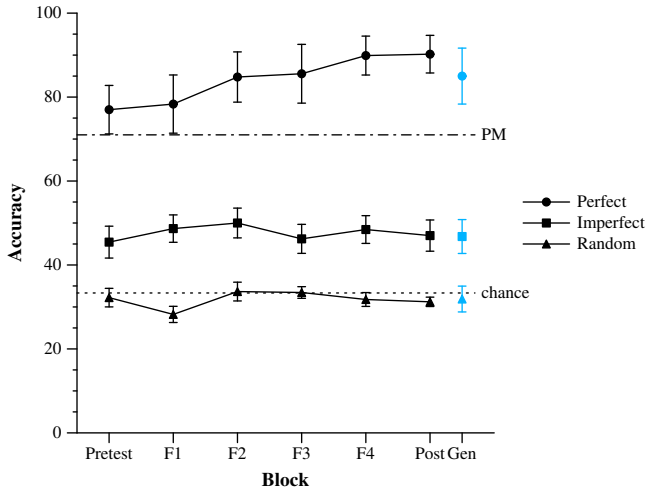

Figure 2.3: Results of Experiment 1, non-mixed conditions. Graphs show the proportion of accurate responses in each block, averaged for all subjects in the non-mixed conditions. Six blocks of observation trials preceded the first plotted point ("pretest"). Blocks labeled F1–F4 used the feedback procedure. The post-test and generalization trials, though plotted separately, were intermixed in the final test section. Error bars show one standard error above and below the mean. The line labeled "PM" shows the Imperfect-condition results that would be expected if subjects were probability-matching; the line labeled "chance" represents the expected accuracy of random guessing.

2.1.4 Results: Accuracy

Non-mixed conditions Figure 2.3 shows the initial results from the Perfect, Imperfect, and Random conditions. The independent variable is the proportion of responses that agree with the mapping relation that was used in constructing the stimuli. Although I call this *accuracy*, remember that in the non-Perfect conditions the subject's feedback may have indicated that the response was wrong. On the trials that were exceptions to the underlying pattern, the subject's feedback indicated whether the subject matched the *exceptional* form, but the accuracy score analyzed here measures whether the subject matched the *underlying* form. Only for the Perfect condition do the two coincide. For the Random condition, what counts as an accurate response is arbitrary, since the underlying pattern was no better a description of the data than any other pattern one might invent. However, for consistency in scoring, responses matching the underlying pattern are still considered correct.

In order to evaluate differences in learning by the subjects, the results from the pretest and post-test were analyzed by means of analyses of variance (ANOVAs). The first was a 3 conditions $\times$ 3 action-dimensions $\times$ 2 test blocks design. Feedback blocks were not included in the analyses presented here, because of the different procedure used for them, but it is clear from figure 2.3 that subjects' responses in the feedback sections were quite comparable to those measured in the test blocks.

All main effects and two interactions—condition $\times$ block and dimension $\times$ block—showed significant differences. The strongest effect was that of condition ($F_{2,24} = 54.1, p < .001$): the Perfect group was most accurate (83% overall), then Imperfect (46%), and finally Random (32%). Specific comparisons confirmed that all groups differed significantly from one another, both in the pretest and the post-test (the smallest difference is Imperfect *vs.* Random in the pretest, for which $F_{1,24} = 5.0, p < .05$)

The effect of block was significant ($F_{1,24} = 6.7, p < .05$), and its interaction with condition ($F_{2,24} = 5.9, p < .01$) showed that the groups followed different learning curves. An analysis of the simple main effects of block for each group showed significant learning over blocks only in the Perfect condition

($F_{1,8} = 14.8, p < .01$). Average scores in that condition improved from 77% in the pretest (range 53–100%) to 89% in the post-test (range 66–100%). Two out of 9 subjects were above 90% in the pretest, and 6/9 in the post-test. Apparently learning the system was quite difficult, however, because the remaining three subjects did not ever reach truly consistent responding despite being shown and trained on perfectly consistent patterns.

The Imperfect group had great difficulty learning the system, with scores always below the probability-matching level (the range from the lowest subject's worst block to the best block of the best subject, including feedback blocks, is 30–71%). There is no improvement over blocks.

Finally, in the Random control group, the responses matched the arbitrary pattern which was called "accurate" about as often as chance would predict, for all subjects and blocks.

The main effect of the action-feature dimension (that is, accuracies of subjects' predictions of direction *vs.* path *vs.* manner of motion) was significant ($F_{2,48} = 4.3, p < .05$), as was its interaction with condition ($F_{4,48} = 2.9, p < .05$). Looking at the effect of the action-dimension in each condition separately, there was only a significant difference in the Imperfect condition ($F_{2,16} = 9.2, p < 0.01$). In this group subjects were more accurate predicting directions (60%) than manners (42%), which in turn were better predicted than paths (38%). This may reflect the differing salience of the three, or the fact that direction and manner were clear from the beginning of motion, whereas indication of the "angle" path (and to a lesser extent, "curve") was delayed until the change in direction could be seen. In the other groups this effect was not significant, probably because of ceiling effects in the Perfect condition, and random guessing for the Random group.

A second ANOVA was required to test whether the object-feature dimensions differed in their ability to be used as predictors, since this variable is entirely confounded with the action-feature used above. An ANOVA identical to the previous one, except substituting object-features for action-features, showed no significant effects involving the new variable.

Subjects also proved to be unperturbed by the generalization trials; accuracy on these was 55% overall, compared with 54% on the trained stimuli. An ANOVA of 3 conditions $\times$ 3 action features $\times$ 2 types (trained and generalization) was run on the post-test data, and confirmed that there were no effects or interactions involving type.

Mixed Condition The data from the Mixed-condition subjects was analyzed separately. Performance graphs are shown in figure 2.4. This graph is analogous to figure 2.3, except that now there are only 9 subjects in all, and each contributed to all three curves.

In general, the average levels of performance are similar to the non-mixed conditions, but variances are larger. This is not only due to the smaller number of subjects, but also to particularly large individual differences in subjects' response to this peculiar condition. There are two subjects who apparently were not able to separate out what was consistent from what was not, and remain at chance on all three dimensions. In contrast, one subject ended up maximizing the imperfect dimension (responding with 100% consistency in the post-test despite the inconsistency of the data in the observation and training blocks). Thus, the mixing of the conditions did appear to lead the subjects to approach each dimension differently than they would have if had been presented as part of one of the non-mixed conditions.

An ANOVA of 2 blocks $\times$ 3 probability-levels (the within-subjects equivalent of the "condition" factor in the non-mixed analysis) showed a significant effect only of probability-level ($F_{2,16} = 11.0, p < .001$). The improvement over blocks was only marginal ($F_{1,8} = 4.7, p < .10$).

2.1.5 Discussion

The mapping relationship turned out to be more difficult for subjects to learn than had been expected, even when they were given perfect information. The majority of subjects fell short of the level of

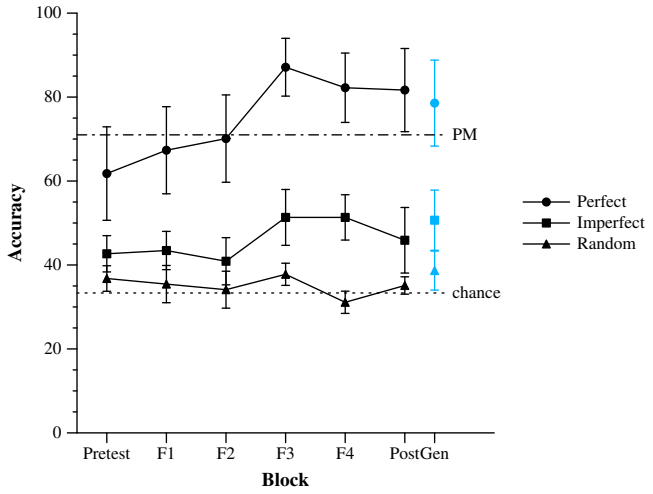

Figure 2.4: Results of Experiment 1, Mixed condition. The proportion of accurate responses over time, for the 9 subjects in this condition. Each subject had one perfect, one imperfect, and one random dimension pair; the three curves show the average over all 9 subjects of the corresponding dimensional accuracies. Error bars show one standard error above and below the mean. The "Post-test" and "Generalization" trials, though plotted separately, were intermixed in the final test section.

probability-matching: their performance curves were bounded between chance and the data's consistency. Since few subjects ever reached the level of probability-matching, it is not possible to answer the question of whether a matching or maximizing strategy would be preferred on this kind of system. The results of this experiment proved useful mainly in suggesting improvements on the paradigm, which are incorporated into Experiments 2 and 3.

First, there were some hints that matching or maximizing might be used if the problem were somehow made simpler. In the current experiment, calculating accuracy for each of the three dimensions separately revealed that several subjects maximized a subpart of the problem: one subject responded according to a shape-to-direction pattern on 100% of the trials, and two with 85% accurate color-to-direction patterns in the post-test. None of these subjects scored above 60% on either of the other dimensions, however.

One way of attempting to find out about the use or non-use of maximizing strategies would be to give subjects more exposure to the systems by continuing the experiment for a longer period of time. However, it is not clear that this would help, since for the non-Perfect groups there is no indication that any learning occurred after the initial set of observation trials. This seemed likely to be due to the confusing nature of inconsistent feedback: the same guess in the same context can be called correct on some trials, and wrong on others. Pilot data and subjects' reports suggested that this was considerably more confusing and frustrating than simply observing inconsistent patterns. Thus, in later experiments the feedback trials were eliminated in favor of increased numbers of observation trials.

The generalization trials demonstrate that those subjects who were successful in finding the patterns did so by finding dimensional mappings, not by memorizing individual stimuli. This is supported by subjects' reports during debriefing after the experiment; most were surprised to be told that there were stimuli in the last block that had not appeared before.

However, the results of the Mixed condition suggest that the dimensions are not considered completely separately; the presence of some randomness in other dimensions apparently caused the perfect dimension to be learned more slowly—and not at all by some subjects. Compare the learning curve for the perfect dimension in figure 2.4 to the learning curve for the Perfect-condition subjects in figure 2.3. The two curves represent subjects' responses to the same rule being learned from the same data, with the

only difference being the context (the other two dimensions) in which that data is presented. However, the resulting levels of accuracy seem quite different.

The fact that the Mixed condition did not show statistically significant differences between learning curves like those found in the non-mixed conditions suggests that the Mixed condition is not a beneficial way of studying learning within this paradigm: the increase in statistical power from using within-subjects comparisons is more than offset by the increased between-subject variability caused by different reactions to the confusing array of stimuli. Although it would be interesting to pursue the effects of mixed probability levels, these specific effects are not of central interest to this project. Thus, it was decided to use only non-mixed conditions in the following experiments.

Experiment 1 thus demonstrates the feasibility of this paradigm for studying the questions of interest, but also suggests some procedural modifications, which were employed in the remaining experiments. Experiment 2 pursues these questions with adults, but employing entirely non-mixed conditions, simpler systems, and longer observational learning with no feedback trials. Experiment 3 employs the same procedures with children.

2.2 Experiment 2

In the second experiment, the search for rule-governed behavior is pursued with simpler systems and simpler presentation. By scaling down the complexity, it was hoped that subjects would easily learn the perfect version of the system, and in the imperfect conditions would absorb enough information to either match the probabilities or form rules.

There were several differences in procedure from Experiment 1, all of which were designed to make the subjects' task easier:

- The system is smaller: subjects were shown objects and actions that varied in one or two (rather than three) dimensions. Thus, there are only 3 ("small" system) or 9 ("medium" system) objects engaging in the same number of actions, as compared to the large 27-object system of Experiment 1.
- Observation trials were used instead of feedback trials, which had proven ineffective in the Imperfect condition of Experiment 1.
- More presentations of each stimulus were used. Overall, subjects in this experiment were given 378 observation trials, as compared to 126 observation trials plus 84 feedback trials in Experiment 1. This is over 5 times as many presentations of each stimulus for the medium system, and over 12 times as many for the small system. The increase was made possible without increasing the total time required for the experiment by the use of observation instead of feedback trials, fewer stimuli, and a slightly briefer stimulus presentation.
- The task was object prediction instead of action prediction, which allowed for much simpler responses: instead of requiring multi-key manipulations of a schematically-represented action, in this experiment subjects simply selected one of an array of response keys, each of which was labeled with a colored shape.
- *Catch trials*, in which subjects were asked to recall the immediately preceding observation trial, were used in order to allow the identification of subjects who were simply not paying attention or didn't understand the task.

2.2.1 Subjects

The subjects were 29 University of Rochester undergraduates (13 male, 16 female, aged 18–22 years), recruited and run in the same manner as the subjects in Experiment 1. No subject participated in both experiments.[1]

2.2.2 Stimuli

The "size" and "path" dimensions were not used in this experiment, since only 9 stimuli were needed and those dimensions had been the least well learned in Experiment 1. The size of all the stimuli in the current experiment is 2cm, and the path of movement always linear.

Two groups of subjects got *medium-sized* systems, with objects varying in color and shape and actions varying in direction and manner of motion. Seven of the 9 possible stimuli were used for training, and two reserved for the generalization tests. The other subjects were given *small* systems, in which the objects and actions had only one dimension of variation: objects varied either only in color or only in shape, and actions either only in direction or only in manner. The three possible stimuli formed the training set. With only one dimension of variation, it was not possible to have generalization stimuli in the small systems.

For counterbalancing, in medium systems, the two object-dimensions were mapped onto the two action-dimensions in both of the possible ways. For small systems, there are four combinations of dimensions possible: shape–direction, shape–manner, color–direction, and color–manner. When color was not used, all stimuli defaulted to color being red; shape, likewise, defaulted to square; direction to left; and manner of motion to smooth (steady sliding) motion. As before, three counterbalanced ways of mapping the values within dimensions were used.

The subjects in the medium systems were either given perfect or imperfect data. The Medium Perfect condition had 100% consistency of mapping between the 2 object dimensions and the 2 action dimensions. The Imperfect condition had an overall consistency of 71%, created by introducing two exceptions on each object dimension in each group of 7 stimuli.

In the small systems, subjects were either given imperfect (71%) or random (33%) mappings between the single object and action dimensions which varied over stimuli. The 71% consistency was created by changing sometimes one and sometimes two of the trials in each set of three, while the Small Random condition was achieved by always changing 2 of the 3.

The trials were again randomly ordered into sequences, independently for each subject and block, subject to two restrictions: the same action could never follow itself, and each of the distinct stimuli had to be presented before any one could be presented again. In the conditions with exceptions, stimuli with identical objects (but not actions) did occasionally follow one another; this was unavoidable because of the second restriction. No restrictions of this sort had been necessary in Experiment 1 because the much larger number of stimuli made sequential repetitions unlikely to happen.

The design of Experiment 2 did not contain the Small Perfect and Medium Random conditions which could have completed a factorial design. The former was deemed unnecessary on the expectation that subjects in the Medium Perfect system would already be at ceiling in their performance. This expectation was confirmed. Since subjects were able to respond accurately on nearly every trial for 7 stimuli, it seems fair to assume they they could have done the same for 3. The Random conditions, conversely, were designed to test for biases that subjects may have in responding to the stimuli, which are unconnected to any patterns in the data. Such response biases should therefore show up regardless of the system presented, but would be presumably be more difficult to find and more variable between subjects if the space of stimuli and possible responses is enlarged. If there are no such biases, then one would expect

[1] One tried.

responses to be distributed evenly across all possible responses, resulting in close to 33% accuracy for all subjects, as was found in Experiment 1.

2.2.3 Procedure

The equipment used was the same as in Experiment 1, except for the response choices. In this experiment 9 response keys were available, each labeled with a single colored object: red square, green square, blue square, red plus, and so forth, arrayed in a 3×3 matrix.

The experiment consisted of alternating observation and test blocks. The observation trials were identical to those in Experiment 1, except that the distance that the stimuli traveled along their trajectory was slightly decreased (the difference was not visually obvious with the linear trajectory, but it decreased the length of each observation trial from approximately 9 seconds to 7).

Also, one of the observation trials in each group of seven (medium systems) or one in every third group of three (small) was designated a *catch trial*. In catch trials, after completing the observation as normal, the subject was unexpectedly presented with a question mark, and had to pick the response key that pictured the object last seen. No feedback was given on catch trials.

In the test trials, subjects were presented with a black question mark, moving with one of the same directions and/or manners of motion that had been used by the observation items. Subjects were told that the question marks were objects of the kind that they had previously seen, but "their shape and color are hidden," and they should "guess which shape and color it was" by pressing one of the response keys.

There were 126 stimuli in each of the three observation blocks (18 groups of 7, or 42 groups of 3),[2] and 54 stimuli in each of 3 test blocks (6 groups of 9, or 8 groups of 3); the subject was allowed a break after each 21 stimuli.

2.2.4 Results: Accuracy

Catch Trials No subject made errors on more than 6 of the 30 catch trials, so no subjects' data were rejected on that basis. The average was 1.4 errors, or 95% correct on catch trials.

Accuracy Figure 2.5 presents mean learning curves for subjects in the various conditions. As expected, subjects in the Medium Perfect condition were easily able to learn which objects were associated with which actions (the overall average on trained trials was 97%). The aggregate scores for the Imperfect conditions approached the probability-matching level, and the Random condition was close to the expected chance responding.

The initial analysis of the accuracy scores was accomplished, similarly to Experiment 1, by analysis of variance methods. First, the overall effects of the condition and block were analyzed with a 4 conditions $\times$ 3 blocks ANOVA; only the main effect of condition ($F_{3,25} = 22.7, p < .001$) was significant. Specific comparisons between the conditions showed that the Medium Perfect condition was significantly higher than either of the Imperfect conditions ($F_{1,25} = 13.3, p < .01$) which in turn were higher than the Small Random condition ($F_{1,25} = 15.4, p < .001$), but the two imperfect conditions were not different from each other ($F < 1$).

[2] An exception was made for the Medium Perfect condition, however, in which it was noted that subjects were already at ceiling in far less than the full 54 observation blocks; these subjects were only required to complete half that number. As should be clear from the data presented here, this could not have affected the results, since they scored nearly perfectly in all tests despite the smaller number of observations.

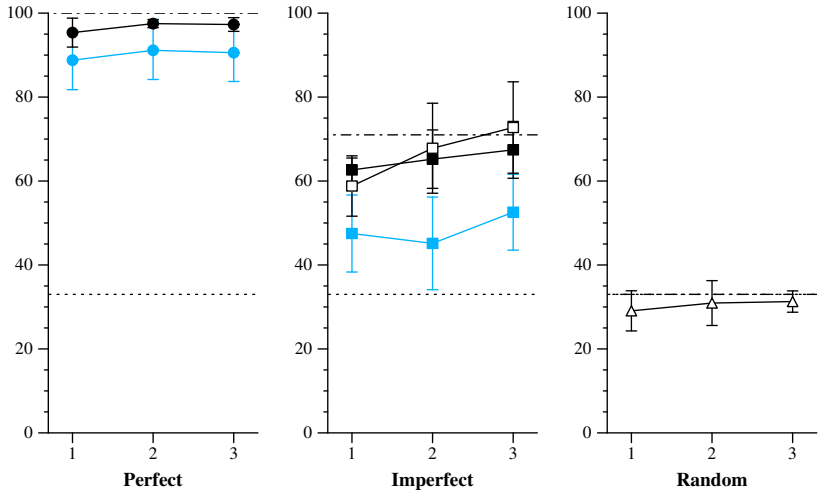

Figure 2.5: Results of Experiment 2. Average accuracy for all subjects by block, for each of the 4 conditions. The solid marks are medium conditions, empty marks are small conditions. The generalization tests (medium conditions only) are shown separately, by the colored curves. For comparison, the dotted horizontal lines show the chance level of accuracy, and dot-dash lines show theoretical probability-matching levels for each level of data quality.

In the last block, the mean accuracies of the imperfect conditions closely approximated the probability-matching level of 71%. However, it may not be accurate to simply state that subjects were probability-matching, since there were large individual differences. In fact, the individual data were strongly reminiscent of Experiment 1, with the majority of subjects' scores confined between chance and the probability-matching level, despite the simplicity of this system.

In the Medium Imperfect condition, the overall average accuracy was 61%. Two subjects (2MI3A and 2MI4C)[3] were more consistent than their data, reaching 98% and 81% accurate responding, respectively. Another two maximized single dimensions: 2MI1B reached 87% accuracy on one dimension (direction–shape), but stayed under 40% on the other (manner–color), while 2MI4C scored 61% on manner–shape and 100% on direction–color in the last block. The remaining 3 subjects for the most part remained in the 40–60% range for all blocks.

In Small Imperfect the overall average was 66%, with even more variability between subjects. Four subjects maximized by the last block, but two remained near chance, with the remaining two falling in between. Thus, despite the coincidence of the group mean with the probability-matching level, it does not look like there is a strong tendency for subjects to probability-match. Rather, most subjects either maximize, or perform somewhat below probability-matching.

The final condition, Small Random, produced some surprises. Three subjects adopted rule-like response strategies in response to the random data. These patterns happened to disagree on each of the stimuli with the arbitrary "correct" response, so they showed up as accuracies of (or near) 0, and affect the overall average little. However, as will become clear later, this turned out to be a very important effect (see the next chapter). The discovery of this phenomenon prompted a re-analysis of the data in terms of consistency rather than accuracy, and a comparison between these subjects and the children tested on a similar system in Experiment 3. This will be taken up again in the next chapter, after the procedures and basic data of Experiment 3 have been presented. As far as accuracy data are concerned, for the majority

[3]The codes used to designate subjects are constructed from the experiment number, the complexity and data quality of the condition, a number designating how the object and action dimensions were mapped onto each other, and a final letter designating how the values within each dimension were mapped.

of Small Random subjects performance was close to the random guessing that was expected; the overall average accuracy was 30%.

Generalization For the Medium conditions only, the generalization trials were compared to trained stimuli with an ANOVA of 2 conditions $\times$ 3 blocks $\times$ 2 types. The condition ($F_{1,12} = 29.0, p < .001$) and type ($F_{1,12} = 7.8, p < .05$) variables both had significant main effects, but their interaction was not significant: trained stimuli were more accurately responded to than generalization stimuli in both the Medium Perfect and Medium Imperfect conditions.

This result may be misleading, however, since inspection of the data reveals that the generalization decrement in the Perfect condition (90% generalization *vs.* 97% trained) was primarily caused by a single subject (2MP4A), whose generalization accuracy was 50%. Excluding this subject, the average for generalization trials rises to 96%, with trained trials remaining at 97% accuracy.

The Medium Imperfect subjects, on the other hand, all scored lower on the generalization than the trained stimuli; average accuracies were 48% and 65% respectively. In this case it is not the result of a small number of subjects; there was some noticeable decrement for every subject in the group.

In Experiment 1, there was no detectable difference between trained and generalization stimuli. However, in the current experiment the combination of a small number of stimuli and imperfect data apparently led subjects to respond based on information about the individual stimuli, rather than about the dimensions of variation.

Subject Variables In order to test whether there were any reliable differences in subjects' scores caused by various characteristics of the subjects, a multiple-regression fit of the mean accuracy scores was attempted using the variables gender, handedness, and age, as well as indicator variables for the conditions. The Small Random condition was not included in the analysis, since the accuracy scores in that condition are arbitrary. Each variable was tested to see if it significantly improved the prediction of accuracy scores, after taking into account the effects of all the other variables.

As was expected, the indicator variable for the Medium Perfect condition was significant ($F_{1,16} = 9.8, p < .01$). No other variable attained significance, although age was marginal ($F_{1,16} = 3.4, p < .10$), with the best fit line showing an increase of 6% in accuracy per year of age.

Each subject had been asked both whether they were left-handed and whether any members of their immediate family were left-handed. Separate analyses were run using these as separate variables, or combined into a single score (nonzero if the subject either was left handed or had left-handed relatives), but had no effect in either case.

2.2.5 Discussion

The results of this experiment, and comparisons with Experiment 1, show that a compositional system of mapping relations can be taught to subjects by means of an observation-test procedure in a small number of trials (most of subjects' learning, in fact, occurred in the first block, which is important for designing a version of this experiment that is short enough to run with children).

Subjects in the Medium Perfect condition performed nearly perfectly, even in the first test block (after only 63 observation trials). Thus, the complexity of the systems used would appear to be easily within the grasp of subjects. This is a strong difference from Experiment 1, where even the Perfect condition proved difficult for many subjects.

However, introducing exceptions into the observation data had a severe effect on accuracy. Of the Medium Imperfect subjects, only one maximized, while the rest remained below the matching level, not following any readily-classifiable response strategy. Even in the simpler Small Imperfect condition,

after 126 exposures to each of just three shapes and actions, only about half the subjects maximized, and several still remained near chance levels of responding.

Thus, many adult learners of these systems do not seem to be able to overlook impoverished data, or even completely accept it as a probability-learning situation. Rather, they appear to follow the basic probabilistic structure, but with a lot of noise or guessing responses diluting the patterns. This is the same pattern that had been seen in Experiment 1; replacing feedback trials with many more observation trials and the other simplifications made did not seem to improve the average very much.

Also, subjects in the imperfect conditions appear to have learned what they learned by memorizing the behavior of particular stimuli, rather than learning generalizations about the dimensions. This is supported by the large difference between the trained and generalization stimuli in the Medium Imperfect condition (and one of the subjects in Medium Perfect).

The most important result of this experiment is the strong effect of data quality on adults' accuracy. There also appears to be an effect of changing the complexity from the very complex Experiment 1 to the more moderate systems used here, but there is no detectable difference between the medium and small systems. Of the three major variables we set out to explore, two—data quality and system complexity—have been considered so far. Thus, Experiment 3 moves on to the third major variable, the age of the learner.

2.3 Experiment 3

This experiment uses the same procedure to compare the responses of first-grade children to those obtained from adults. The stability of adults' performance—as shown by the lack of significant learning after the first block in any condition of Experiment 2—was taken to be an encouraging sign that the paradigm could be useful, and results roughly comparable, even with the smaller number of trials possible to include within the attention span of a 7-year-old. The methods and stimuli here are very similar to those used in Experiment 2, except for the length of the experimental session and some minor changes designed to make the most of the shorter time available.

In response to some surprising data that came out of the first few days of testing, several additional conditions were run in this experiment. Thus, some of the conditions to be described are near replications of Experiment 2, while others have not been tried with adults.

2.3.1 Subjects

The subjects were 69 children (38 male, 31 female) enrolled in first grade classrooms of the Brookview and Listwood schools in Irondequoit, NY—both public elementary schools in a middle-class suburban school district. The median age was 7 years, 2 months (range 6;6–8;9). Two additional subjects participated in the study, but were not able to complete the experiment due to a scheduling error.

2.3.2 Stimuli

The stimuli were the same as those used in Experiment 2. The conditions used were Small Perfect, Medium Perfect, Small Near-Perfect (85% consistent data), Small Imperfect (71%), Medium Imperfect (71%), Small Random, and Medium Random. A Mixed condition (100%/71%) was started, but not enough subjects were available to complete it.

Dimensions and values were counterbalanced as in Experiment 2, but the restrictions on the random order were changed: in this experiment the overriding rule was that no two stimuli whose objects *or* actions were identical could follow one another.

2.3.3 Procedure

The equipment setup was the same as in Experiment 2, but an additional computer was used so that two subjects could be run at once; this was a Macintosh LC when running at one school and a Macintosh LC575 at the other. The program's timing was adjusted to compensate for the different processor speeds of the three machines, so that trials would take the same amount of time for all subjects.

Subjects were run two at a time in a quiet room in their school. They were seated at the computers such that they could see only their own screens, but the pairs were given instructions together and ran simultaneously. The experimenter remained in the room, seated between the subjects, and started the two computers at the beginning of each block. After completing the experiment each subject received a small reward (a drawing and a sticker).

As in Experiment 2, the experiment consisted of 3 observation blocks alternating with 3 test blocks, but in this experiment the observation blocks contained only 42 trials, and the test blocks 27 trials each. The number of observation trials in the entire experiment was thus equal to the number of trials in the first observation block of Experiment 2, and each of the test blocks was half the length of one of Experiment 2's test blocks. Since each block was so short, no breaks were needed except between blocks. The entire experiment lasted approximately 20 minutes.

2.3.4 Results: Accuracy

Catch trials Subjects in the small conditions averaged 3.7 errors out of 15 catch trials (75% correct), while medium subjects averaged 4.8 errors out of their 19 catch trials (also 75% correct). Seven subjects got 1/3 or less of the catch trials correct, which is chance performance (two in small conditions, with 10 errors each, and five in medium conditions with 13 to 16 errors). Their data were not included in the analyses, and other subjects were run in their place, since it was assumed that they either had not understood the procedure or were not attending to the experiment. The remaining subjects average catch-trial accuracies were 78% (small) and 84% (medium).

Subsequent inspection of the discarded data, however, did not show any obvious systematic differences between the subjects who missed many catch trials and those who did not. Average accuracy of the replaced subjects in the small conditions seemed lower than the overall average (31% *vs.* 46%), but there was no correlation, including all groups of subjects, however, between number of catch-trial errors and accuracy scores ($r = .04$).

Accuracy Figure 2.6 presents the learning curves for subjects in the seven non-mixed conditions of Experiment 3. Overall, their level of mastery of the systems presented was considerably lower than adults given the same systems. In fact, the children's overall scores on the small, 3-stimulus systems are similar to the adults' means for the largest system (the 27-stimulus system of Experiment 1). Initial analyses showed overall averages for both Perfect groups to be significantly above chance (Medium: $t_7 = 3.3, p < .05$; Small: $t_7 = 3.3, p < .05$), and the Small Imperfect group marginally above chance ($t_7 = 2.1, p < .10$). The Small Random group was marginally *below* chance ($t_7 = -1.9, p < .10$); since the response scored as correct was arbitrary for the Random groups, this can be taken simply as an indication of consistent responding, and as such will be further explored in chapter 3.

Ignoring for now the Near-Perfect condition, the other six groups were analyzed as a factorial design crossing complexity and data quality. Thus, an ANOVA of 2 system sizes $\times$ 3 levels of data quality $\times$ 3 blocks was run. The main effect of data quality ($F_{2,42} = 12.3, p < .001$) and the interaction of data quality with block ($F_{4,84} = 2.7, p < .05$) were the only significant effects. The interaction is due to an improvement in accuracy scores over blocks for the Perfect conditions, and the lack of improvement in the Imperfect and Random conditions (the simple main effect of block for Perfect conditions only is $F_{2,28} = 4.7, p < .05$; for Imperfect and Random, $F < 1$). Accuracy in the Perfect conditions was higher

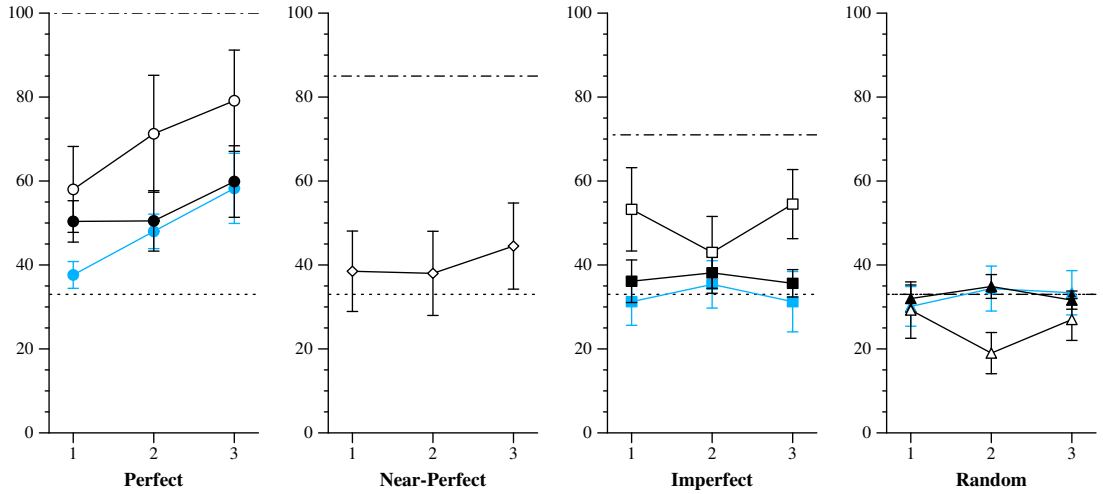

Figure 2.6: Results of Experiment 3. Average accuracy for all subjects by condition, over blocks. The graphs show diminishing levels of data quality; within each, the outlined marks are small systems and the filled marks are medium systems. The colored curves are the results on generalization trials.

than in Imperfect for blocks 2 ($F_{1,42} = 7.1, p < .05$) and 3 ($F = 11.4, p < .01$). The difference between the Imperfect and Random conditions was significant in block 3 only ($F_{1,42} = 4.2, p < .05$).

The remaining condition, Small Near-Perfect, was added in hopes of getting higher levels of accurate responding than were seen in the Imperfect systems. This group's observation trials were 85% consistent, putting the condition midway in quality between the Imperfect and Perfect conditions. However, the group's mean accuracy was 39%, statistically equal to the other non-perfect conditions (specific comparisons of the means within an ANOVA containing all 7 conditions of the experiment $\times$ 3 blocks, showed Small Near-perfect to be significantly different only from Small Perfect, $F_{1,49} = 9.6, p < .01$). Also, like the other Imperfect conditions, Near-perfect showed no significant change over blocks. This suggests that once there is inconsistency in a system, the number of exceptions is not particularly important to children—even 15% inconsistency leads to major effects on learning.

Generalization In parallel with the analysis of Experiment 2, a data quality $\times$ block $\times$ type ANOVA was run on the medium systems to determine whether there was a difference between the generalization trials and the trained stimuli. In contrast to the adults, however, children showed no effect or interactions involving type; generalization trials caused no extra difficulty for the children.

Subject Variables As in Experiment 2, a multiple-regression analysis was performed to see if various characteristics of the subjects were affecting the results. The factors in the regression were: gender, familial handedness, age, which computer was used, and indicator variables for the conditions. Only the Small Perfect condition indicator ($F_{1,27} = 7.5, p < .05$) and gender ($F_{1,27} = 4.7, p < .05$) had significant effects. The best fit estimates the boys' accuracy scores 16% higher than the girls'.

Age, which had been a significant predictor in Experiment 2, was not a reliable predictor of performance here ($F_{1,27} = 2.7, n.s.$), but the trend was in the same direction (accuracy scores increasing with age).

2.3.5 Discussion

The experimental procedure proved to be quite easy to use with 7-year-olds. They were, as might be expected, more easily distracted, but also turned out to be more enthusiastic about participating. Despite the fact that several subjects had to be eliminated because of extreme numbers of catch-trial errors, in general the children seemed clear on the instructions and comfortable with the procedure. The data clearly show that many subjects, especially (but not exclusively) those in the small conditions, learned a significant amount from the data they were shown, despite the short amount of time available in this version of the experiment.

There were also some quite surprising aspects of the data. Accuracy in several of the conditions was very low—even below chance—and in looking for possible explanations for this, we found that some of the children who were responding most consistently were following patterns that were not those in the observation data.

Before taking up this interesting finding, however, we summarize the results of all the accuracy comparisons. Since the major interest of this experiment was how the children would compare to adults, the discussion offered here will involve the data from all of the experiments, and in particular Experiments 2 and 3, which directly compared adults and children on the same task.

2.4 General Discussion of Accuracy Results

At the outset of the experiments, three major variables (data quality, system complexity, and age of learners) were chosen for investigation. These variables were chosen as vehicles to test the more general question of whether the system would be learned in a "language-like" way. In general, the accuracy data described here do not support the idea that the learning going on in this experiment is language-like. Rather, the results are in line with expectations if one considers this experiment to be a complex analytic problem-solving task—a particular kind of task that children have great difficulty with, but which college students are routinely faced with and expected to master.

Data Quality The primary variable whose effects we were interested in was the quality of the data from which the subjects were supposed to learn. This variable was the only one manipulated within all three experiments, and it had powerful effects in each one.

The perfect systems were readily learned by all groups—every group either reached near-perfect responding, or was still improving in the last block. Thus we have no reason to doubt that all groups could have solved these problems if they were given sufficient time.[4]

As soon as exceptions were introduced into the data, however, accuracy dropped off dramatically. Although there were some subjects in each experiment who maximized (or maximized on some of the dimensions), performance in the imperfect (and near-perfect) conditions was characterized overall by accuracies between the chance and probability-matching levels. It is worth stressing that the low accuracy scores in the imperfect conditions do not appear to be effects of the systems simply being too difficult to learn. The same systems were learned well by subjects given perfect information. Also, the low scores are found for systems as small as 3 stimuli, and (at least for children) with as few as 15% exceptional trials. Furthermore, no improvement over blocks was found for any Imperfect group. All these results suggest that it is not simply a case of the task getting slightly more difficult; rather, there is a stable, qualitative difference between learning from perfect and imperfect information. Some implications of this will be considered in chapter 4.

[4] A possible exception is in the Mixed condition of Experiment 1, where it is not clear that accuracies were continuing to improve. However, this is presumably not due to the difficulty of learning from perfect information *per se*, but rather due to the presence of exceptions in the other parts of the system they were learning.

Complexity One major reason the complexity variable was included was in order to test a prediction of Newport's Less Is More hypothesis—that the maximizing seen in language learning could be a consequence of a cognitively-limited learner trying to make sense out of a complex system. While the complexity of the systems used in this experiment is much less than that of many structures in languages, a trend towards increased maximizing with increasing complexity would tend to support the hypothesis as an explanation of those age-related changes in language learning.

However, complexity had surprisingly little effect on learning in these experiments, and what effect there was was in the opposite direction. There were no noticeable differences between the Small and Medium Imperfect systems (which had 3 and 9 stimuli) compared in Experiment 2. Likewise, children showed no reliable effect of system complexity; the non-significant trend was toward higher accuracy on simpler systems.

The only clear complexity-related difference is between adults' accuracies on the medium systems of Experiment 2 and the large (27-stimulus) system used in Experiment 1. However, there were also significant procedural differences between the two experiments—the response procedure in Experiment 1 was more cumbersome and confusing, and the feedback trials were apparently not as useful as the extra observation trials which replaced them. Thus, it is not clear that we can interpret this difference as due solely to the difference in complexity of the systems.

Therefore we find no support here for the prediction that maximizing should increase with increasing complexity. While the complexity of these systems is far from the complexity of natural languages, these data do not suggest that the rule-governed nature of language learning arises simply from the sheer complexity of the system to be learned. It is still possible that complexity is one of a complex of factors that are involved in this effect, but if so, then this experiment failed to include some of those critical features.

Age Age is known to affect language learning in various ways. Of particular interest here is the interaction between data quality and the age of the learner: in language acquisition, children are relatively insensitive to the quality of the data they have to learn from, ignoring unpredictable exceptions in favor of forming rules. Adults, however, are more likely to closely match their language model, even if that model uses linguistic structures inconsistently.

Thus, if this system is learned in a language-like way, we would predict more probability-matching in adults, and more maximizing in children. One place this could show up is in the difference between accuracies in the Perfect and Imperfect conditions; the prediction would be that this difference would be large for adults, but small for children.

Adults in general learned all the systems more quickly and successfully.[5] However, for both adults and children performance on the Imperfect conditions was at about 70% of the level of performance on the Perfect conditions; this is more consistent with a probability-matching interpretation than with the prediction of maximizing in children, although the large individual differences also shed some doubt on whether probability-matching is the proper description of the data: for most subjects, inconsistent data seemed to produce somewhat worse performance than probability matching.

Overall, the age comparison yielded results that simply reflect the fact that adults are more skilled at general problem-solving tasks than children. At least from the point of view of accuracy of learning, there is no indication that the particular task under investigation is, like language-learning, one at which children are particularly adept.

[5] One possible, but unproven exception is the response to the generalization test. Adults in the Medium Imperfect condition did significantly more poorly on generalization trials than on the trained stimuli. In contrast, there was no generalization decrement found for any group of children, but the direct comparison to the children's Medium Imperfect condition is unfortunately meaningless: the children scored too close to chance for a generalization decrement to be noticeable. Thus, we cannot say for sure if children were more apt to accurately generalize than adults.

Summary In general, the accuracy data are consistent with the prior literature on the learning of complex systems. Accuracy is best for adult learners who are given perfect data, and particularly for those asked to learn very simple systems. The various tests for language-like learning patterns yielded negative results.

One explanation for these results would be that one simply should never expect language-like learning of a non-linguistic system. The algorithms or neural mechanisms used for learning in the linguistic domain may be entirely different from those used in other domains, leading to different performance on the different systems.

A second possibility, however, is that the structure of the systems used here was simply not sufficiently language-like. Perhaps a still more complex system, or one which shared some crucial property of languages that was overlooked in the design of these systems, would reveal language-like learning.

Before deciding between these possibilities, however, it is worthwhile to take another look at the data. As has been noted several times before, the analyses presented so far have examined *accuracy*—that is, the tendency for subjects to acquire the mapping patterns represented in the observation stimuli. However, as has also been noted, a closer look at the individual subjects' data reveals that many of the children were responding with a high degree of consistency, following patterns which were *not* the dominant patterns of the observation stimuli; the patterns they were following were their own innovations. Indeed, this kind of innovation was found to be rampant in the children's data, and also showed up to some degree among the adults. Since this is reminiscent of some phenomena in language learning, where children fit patterns that seem to be their own innovations into the framework of a system they are learning, the next chapter is devoted to a reanalysis of the data uncovering these innovated patterns, comparisons of their use by adults and children, and some speculations on the relevance of this discovery to the language metaphor.

3 Further Analyses: Consistency and Innovation

The previous chapter presented results of the experiments in terms of *accuracy*, which was defined as the proportion of trials on which subjects responded according to the pattern that had been most common in the observation data—or, in the case of the random conditions where there was no dominant pattern, responses were scored as accurate if they conformed to a particular pattern arbitrarily chosen by the experimenter.

The data discussed in the present chapter are derived from different metrics, based on the idea of looking for consistency in subjects' responses regardless of whether they followed the patterns that were suggested in the observation trials or not. The impetus for these analyses came from the discovery, in the course of analyzing the accuracy data, that some subjects in the random conditions were scoring significantly below chance, some even achieving accuracy scores of 0. This is indicative of great consistency, but according to a rule that disagrees with the pattern called "accurate." An example of this can be seen in figure 3.1, which shows the responses of one child in the Small Random condition of Experiment 3. This subject was given observation trials where colored squares moved in the three directions, with no correlation between the colors and directions. Then, in the test trials summarized in the figure, she watched a question mark move in one of three directions, and was asked to guess the color of the the object that moved that way. For each of the three directions, this subject chose a single color on nearly every trial.

Upon further analysis, it turned out that this was not an isolated example, but in fact this kind of pattern-following was very common, particularly among children. Many of the inaccurate responses of children in the perfect and imperfect conditions, in fact, fell into very consistent patterns of responding. As will be shown, children's behavior can be described as tending to follow some pattern consistently, whether or not that pattern is accurate.

This innovative pattern-following behavior is considered here in detail, since it has interesting parallels to some phenomena in language acquisition. Children learning languages are known not to simply mimic what they hear, but to productively use rules to create new utterances, as a number of the phenomena of language learning discussed in section 1.2 attest. Overregularization errors are created by using a learned rule in an improper context, combining morphemes that the child, presumably, has never heard together. Simon, the subject of Singleton and Newport's study, was also found to have innovated combinations of forms which his parents used only separately (see Singleton, 1989)(Singleton, 1989). Furthermore, in the process of creating a creole language, children may innovate entirely new forms and structures (Sankoff, 1979).

Innovation thus appears to be a characteristic property of children's language acquisition. The apparent innovation in children's learning of the experimental systems is therefore worth a careful look. In the context of the experiment, the only way an innovative pattern (as opposed to a simple mistake) can be identified is if that same pattern continues to be followed in a consistent manner. Thus, we begin by defining a measure of the consistency of a subject's responses that is independent of the particular pattern that the subject chose to follow, and present summaries of the consistency of the various groups of

Stimulus Direction:	Response: red	green	blue
↙		2	**25**
→		**25**	2
↖	**26**		1

Table 3.1: Responses of one child to the Small Random system. The table shows the number of times each of the available responses was chosen in answer to each of the stimuli by subject "3SR3C." Overall, 76 out of 81 responses are seen to follow her favored pattern (shown in bold), for a consistency score of 93%. Note that the *accurate* responses, by arbitrary decision, would have fallen on the opposite diagonal.

subjects. Following this, we look for the sources of the consistency, which turn out to be quite different for adults and children.

3.1 Consistency

There are many different kinds of patterns that a subject could choose to follow. For example, responding with a repetitive sequence of responses such as "red, green, blue, red, green, blue…" would be a kind of regularity, although not one that bears any resemblance to the form of the data in the experiment or that has any particular linguistic interest. Such *keyboard patterns* are commonly found in probability-learning experiments, especially with children (Weir, 1964; Estes, 1972). However, such patterning did not appear to be a common strategy by subjects in this study; although some subjects did use them, overall they were not much more common than chance would predict and their use did not differ appreciably between groups or even between ages. Thus this kind of patterning is not considered further.

More interesting than the simple keyboard patterns are patterns of the type that were actually used in the experiment—associations between features of the objects and actions (such as "bouncing means red"). Among the various kinds of patterns that we looked for, this appears to be the only kind that was used by a large number of subjects.

Consistency Defined In order to measure this kind of patterning, we looked for the shape or color that each subject most often chose in response to each direction or manner of motion. Once the most common response was known, the percentage of trials was calculated on which that most-common response was chosen. This is the consistency for that feature. The feature-by-feature consistencies were averaged to get the subject's overall consistency. For example, in figure 3.1, the boldface numbers indicate the responses that were considered consistent for each of the stimuli. The consistency score is simply the proportion of responses that fall into those boldfaced cells of the matrix—93% in this case.

For small systems, since each stimulus has only one feature (direction or a manner of motion), as does each response (a shape or a color), it is simple to calculate the consistency of a stimulus directly: the consistency for stimulus i is

$$\mathcal{C}_i = \frac{1}{N_i} \max\left(R_{i1}, R_{i2}, R_{i3}\right) \tag{3.1}$$

where R_{ij} is the number of times the subject chose response j to stimulus i, and $N_i = \sum_j R_{ij}$ is the number of occurrences of the stimulus in the period of time under consideration (here $N_i = 9$; see below). The overall consistency, $\mathcal{C}$, is the mean of the $\mathcal{C}_i$'s.

		Response:								
		■	✚	★	■	✚	★	■	✚	★
Stimulus Direction:	↙		**5**	4			**9**	**8**		1
	→	1	3	**5**	**9**				**9**	
	↖	**9**				**9**				**9**
		Test 1			Test 2			Test 3		
		$\mathcal{A} = 11, \mathcal{C} = 70$			$\mathcal{A} = 0, \mathcal{C} = 100$			$\mathcal{A} = 96, \mathcal{C} = 96$		

Table 3.2: An example of innovative responding. The tables show the number of each type of response to each stimulus by one child in the Small Near-Perfect condition (subject 3SN1C); below each table the accuracy and consistency for the block is shown. 85% of the data that this subject saw during the observation trials followed the same diagonal pattern that he finally adopted in the third block. The patterns used in the first two blocks were innovated by the subject.

For Experiment 3, consistency was calculated separately for each 27-trial test segment. This is because it was noted that some subjects changed patterns from one block to the next, but remained consistent within a block (see, for example, figure 3.2). For symmetry, consistencies for Experiment 2 were also calculated for 27-trial segments by dividing each of the 54-trial test blocks in half, and calculating consistency separately for each half-block. Therefore, in all the consistency calculations reported here $N_i = 9$, since there were 9 occurrences of each stimulus in each 27-trial segment.

For Medium systems, the consistency calculation is slightly more complex. Both of the dimensions of variation of the objects and actions have to be considered; subjects are guessing both the shape and color of the test objects, and both of those guesses have to be accounted for. Each of those guesses, in turn, could be based on either the direction or the manner of motion of the stimulus,[1] regardless of which of those mappings the target pattern used. $\mathcal{C}$ values, as defined above, were therefore calculated for both possible pairs of mappings: one score was formed by averaging the $\mathcal{C}$'s for direction-to-shape and manner-to-color mappings, and another score was formed by averaging $\mathcal{C}$ values of manner-to-shape and direction-to-color mappings. The higher of the two averages was used as the subject's consistency. This maximizing operation, designed to make sure that no pattern that subjects might use would be overlooked, does make it likely that consistency scores on medium systems will be larger, just according to probability, than consistency scores on small systems.

Note that consistency, as defined, must always be greater than or equal to 33%. A consistency of 33% indicates that a subject has used each of the 3 possible responses equally often in response to each of the stimuli, so that each of them gets 1/3 of the total. Any other arrangement of responses will favor one of the stimuli, and have an higher consistency value. For this reason, all consistency graphs to be presented are plotted with the Y axis beginning at 33%.

If a subject were just responding randomly, by chance one would not expect their responses to fall into a perfectly even distribution. For a particular number of trials and possible responses, it is straightforward to enumerate the possible ways in which the responses could fall, and thus calculate the expected value of consistency. For the 27-trial blocks with 3 stimuli and 3 responses which were used here, this chance value of consistency is 50%.[2]

[1] Or, in theory, both. There is little evidence that any subject used this more complex type of rule, however.

[2] Calculated by the counting rule for the 9 trials and 3 responses that were made to each stimulus. Averaging over stimuli does not affect the expected value, although it does narrow the variance of the distribution. For medium systems, the expected consistency is slightly higher than 50%, because of the operation of taking the maximum of the two possible dimensional mappings, and the variance somewhat lower because of the extra averaging.

Entropy Other measures of the consistency of responses are also possible. For instance, in information theory, a standard measure of the disorder in a partitioned set is its *Entropy*. The entropy of the set of choices that a subject makes could serve the same purpose as the consistency measure described here, although with the opposite sense: entropy values are lowest when consistency is highest. Entropy values can range from 0 (perfect order, or 100% consistency) to $\log N$ (total disorder; 33% consistency). For comparison with equation 3.1, the entropy value for a subject's responses to stimulus i would be given by

$$H_i = \sum_j \frac{R_{ij}}{N_i} \log\Big(\frac{N_i}{R_{ij}}\Big). \tag{3.2}$$

The overall measure of entropy corresponding to our $\mathcal{C}$ would be the conditional entropy of the subject's response given the stimulus, which works out to be simply the mean of the H_i's (Baclawski, Rota & Billey, 1989).

The advantage of using entropy rather than the consistency measure defined above would be that the number of responses in every cell is taken into account, rather than just the number of responses in the maximal cell. Say subject A chose the three possible responses to a particular stimulus 5,2, and 2 times, respectively, and subject B chose them 5,4, and 0 times. Both subjects have a consistency on that stimulus of 5/9; but the entropy of subject A's responses is greater than subject B's, since A's distribution of responses is more spread out.

On the other hand, the consistency measure has several advantages over entropy:

- It is simple to calculate, and its characteristics are easily understood.
- Consistency scores are directly comparable to accuracy scores; both are expressed as the percentage of trials on which responses followed a certain pattern (averaged over dimensions, for medium systems).
- The *difference* between consistency and accuracy is a meaningful quantity: it is the percentage of trials (averaged over dimensions) which are responded to according to a consistent pattern which is *not* a pattern from the observation data. This quantity, which I call *innovation*, will be discussed further in section 3.3.

To see if any interesting data would be lost as a consequence of using consistency rather than the more sensitive entropy measure, entropy scores were calculated for all subjects' data, and plotted against consistency scores. No major deviations from the expected logarithmic curve were found. Therefore, the more versatile consistency variable was used for all the analyses that will be presented here.

Consistency Data Figures 3.1 and 3.2 show the average consistency scores for each of the conditions in Experiments 2 and 3. There are several striking features of the data. For the adults, the consistency graphs are very similar to the accuracy graphs presented in figure 2.5; consistency is highest when data quality is highest and, with the exception of the Small Random condition, is roughly equal to the group's accuracy. For children, however, the most striking feature is the evenness of the consistency scores. Unlike accuracy, the level of consistency is close to constant for children, both for particular groups of subjects over time, and between the groups at different levels of data quality.

ANOVAs were run comparing the consistency scores in each of the two experiments, parallel to the accuracy analyses described in the previous chapter. First, a 4 conditions $\times$ 3 blocks ANOVA was run for the adult data from Experiment 2. A significant effect of condition ($F_{3,25} = 5.6, p < .01$) was caused by the high consistency in the Medium Perfect condition (specific comparison of Medium Perfect with Small Imperfect: $F_{1,25} = 6.9, p < .05$); the differences between the two imperfect conditions and Small Random were not significant.

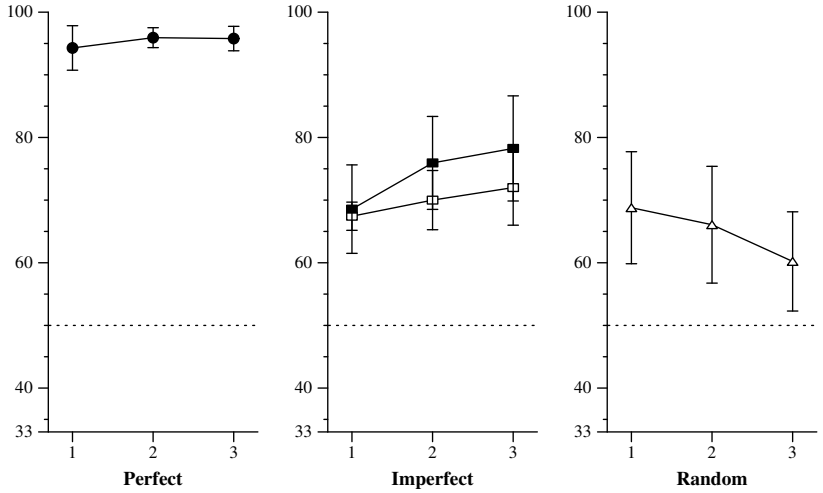

Figure 3.1: Consistency scores for Experiment 2 (adults). See explanation below.

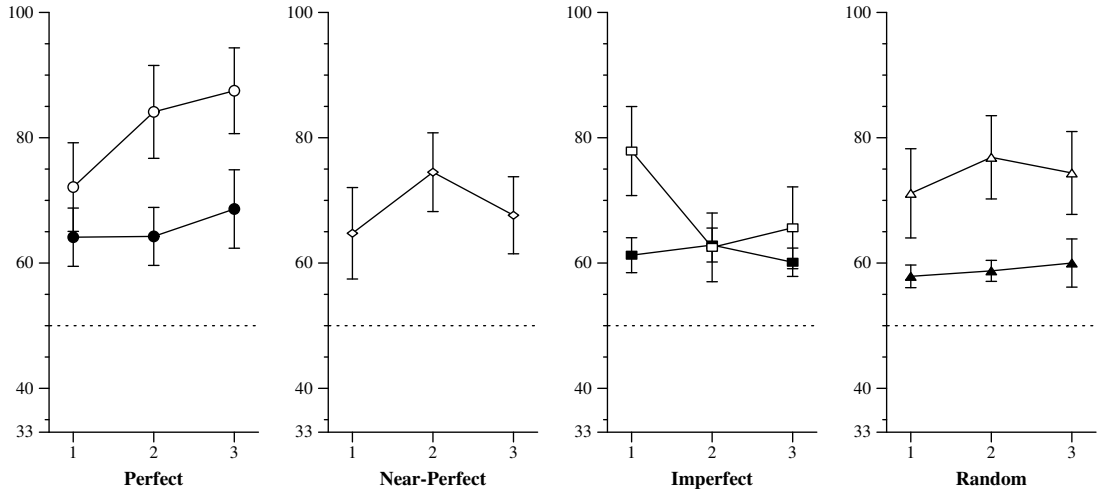

Figure 3.2: Consistency scores for Experiment 3 (children). Within each probability level, the filled marks show average scores for the medium group, and the outline marks are for the small group. Note that the Y-axes begin at 33%, which is the minimum possible consistency score. The horizontal dotted line shows the consistency score to be expected for a randomly-guessing subject.

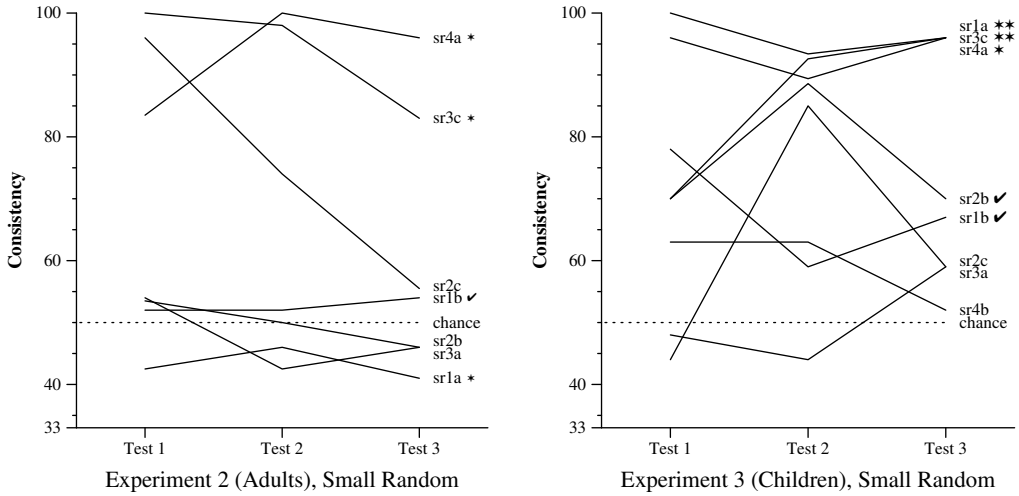

Figure 3.3: Consistent responding by individual subjects in the Small Random conditions of Experiments 2 and 3. Most children show some amount of pattern-following, while most adults respond randomly. A few adults, however, look more like the children according to this measure. Each subject's three blockwise consistency scores were compared to chance (50%) by two-tailed t-tests; ✔, ✶, and ✶✶ designate $p < .10$, $p < .05$, and $p < .01$ respectively.

In contrast, for children, data quality did not have a significant effect. An ANOVA comparing 3 levels of data quality $\times$ 2 levels of complexity $\times$ 3 blocks showed an effect of complexity ($F_{1,42} = 11.3, p < .01$), with small systems scoring an average of 13% greater consistency than medium systems, and an interaction between data quality and block ($F_{4,84} = 3.2, p < .05$). The three-way interaction was marginally significant ($F_{4,84} = 2.4, p < .10$). The locus of these interactions appears to be the increase in consistency over time in the perfect conditions, which unlike the imperfect and random conditions, had significant differences between blocks ($F_{1,14} = 4.2, p < .05$).

Thus, children seem to respond with a given level of regularity regardless of the quality of data they get about a system. Unlike adults, who respond most consistently to systems that are presented most consistently, children's consistency seems to be independently produced.

Although data quality does not seem to affect children's consistency, the system's complexity does. In all cases children were more consistent on small than medium systems. One could speculate that this difference reflects the fact that in order to respond consistently to a medium system, subjects would have to be following—and either learning or innovating—twice as many rules as they would for a small system. Adults did not show any difference, just as they did not show a complexity-based difference in accuracy.

3.2 Sources of Consistency

One problem with the consistency measure is the confounding of pattern *following* with pattern *invention*. Accurate responses are scored as consistent, just like responses that follow any other pattern, but we would like to be able to see to what extent subjects are following the patterns of the observation data versus following patterns that are their own inventions.

Random Conditions One case in which we can avoid this ambiguity is by looking at consistency in the random conditions. In these conditions, there were no patterns in the observation data, so that any

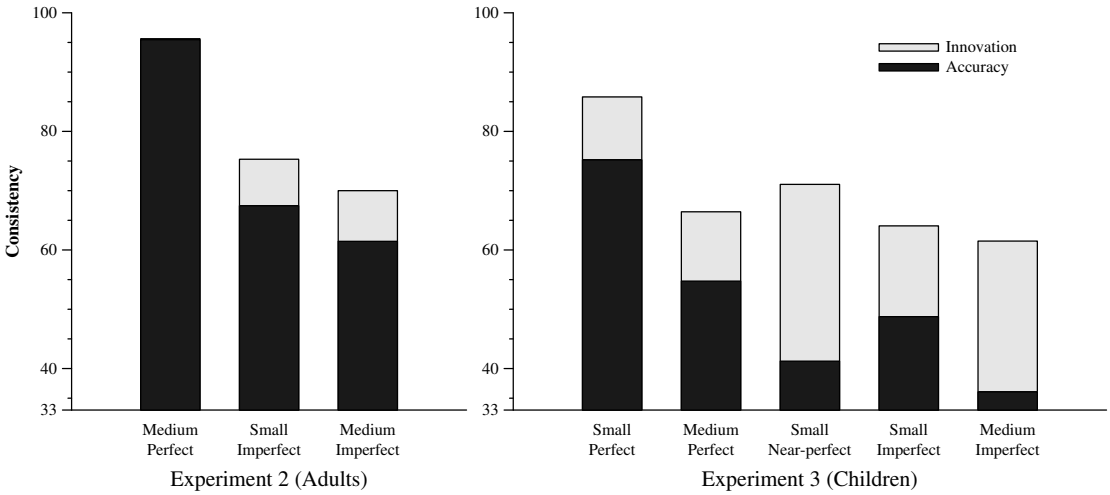

Figure 3.4: Sources of consistent responding for adults and children, non-random conditions. The dark bottom portion of each bar is the average accuracy of the group; the total height is their consistency. The lightly shaded portion, which is called *innovation*, is that portion of subjects' consistency that is derived from following patterns invented by the subject.

above-chance consistency must have been invented by the subject. Both children and adults were found to show some consistent responding in random conditions (figures 3.1 and 3.2). However, figure 3.3, which details the Small Random condition data for each subject, shows that three of the adult subjects are responsible for all of the consistency found in their group. Most of the adults, in fact, responded with what appears to be random guessing; their consistency is just what one would expect by chance. Two subjects were quite consistently following innovated patterns in their answers, and a third began consistent but declined to random responding by the last block. Finally, one subject distributed her responses significantly *less* consistently than chance, which is to say, with a more spread out distribution than random guessing would produce.

For the children, in contrast, consistent responding was the rule rather than the exception. The majority of subjects had consistency scores above chance. Despite the patterns in the observation data to which they were exposed and on which they were told to base their answers, children seemed quite readily to adopt the solution of answering based on a pattern that they had made up. As we will see, the same is true in the other conditions of the experiment.

Non-Random Conditions In the conditions in which there are patterns to follow, nearly all of adults' consistency can be accounted for by accurate responding. Since a subject who is responding accurately is of necessity also responding consistently, it is not possible for subjects' consistency score to be less than their accuracy score. It is, of course, possible for consistency to be *higher:* the difference between the two represents the extent to which the subject following an innovated pattern.

Figure 3.4 shows how much of each group's consistency is the result of accurate responding, and how much the result of innovation. In order to provide a fair comparison between children and adults, the data in this figure comes from selected blocks of the two experiments. Since the adults' observation blocks were three times as long as the children's, and the test blocks twice as long, selecting only block 1 of Experiment 2 and comparing it with blocks 2 and 3 of Experiment 3 is appropriate. Those blocks contain 54 test trials of each experiment, the last of which were after 126 observation trials in each case.[3]

[3]The difference still remains that the adults received all 126 observation trials before the first test block that is included in the

For adults in all the non-random conditions, the difference between accuracy and consistency is quite small. What differences there are, in the Medium Perfect and Small Imperfect conditions, were primarily due to a small number of subjects in each condition, with the majority of subjects almost never consistently follow inaccurate patterns. This is the same pattern that was found in the Small Random data discussed above.

In contrast, children's consistent responses are much less often the result of accuracy or single exceptional subjects, and more often the result of widespread non-accurate pattern following. In fact, the surface similarity of adults' and children's consistency scores (which averaged 76% and 68%, respectively) now appears to be quite a surprising coincidence, since children's accuracies are so much lower. The children make up this difference by creating patterns, and then following those invented patterns with high consistency.

Thus, although children and adults both display high consistency, this consistency came primarily from different sources. Adults are consistent because they follow the patterns of the data shown to them; their consistency is directly influenced by the quality of that data, and, except for some isolated exceptions, if they have not "figured out" the patterns in the data they will guess randomly. Children, on the other hand, are equally consistent on all the systems of a given complexity, regardless of the quality of the data they get about the system. Some of this consistency may be accurate, depending on the difficulty of the system, but only when accuracy exceeds the baseline 60-75% consistency does it begin to show any effect on overall consistency of responses.

3.3 Innovation

The difference between a subject's consistency and accuracy scores has been termed *innovation*, representing the tendency of the subject to create and follow patterns that had no antecedent in the observation data. It has been suggested that children are more prolific innovators than adults, an assertion that is tested in this section.

Although innovation has been described only in terms of the non-random conditions, it would be helpful to have an innovation measure that was applicable to all subjects. In random conditions, however, there is the complication that the accuracy scores are arbitrary. Consider two hypothetical subjects, both with 100% consistent responses, whose accuracy scores are 0% and 100% respectively. If the two subjects were in a perfect condition, the one with 100% accuracy would clearly be following the pattern of the observation data, and would thus have an innovation score of 0; while the one with 0% accuracy, consistently violating the pattern of the observation data, would be credited with 100% innovation. If the two subjects were in a random condition, however, then the only difference between them is whether the pattern they followed happened to coincide with the arbitrary pattern used to score accuracy—a pattern that they had no way of knowing. The consistent pattern of both subjects' responses must have been invented by the subjects themselves. Thus, these two hypothetical subjects should receive the same innovation score: the amount of consistency in their responding that cannot be accounted for by patterns in the data. The observation data shown to random-condition subjects follows each possible pattern 33% of the time, so both subjects ought to be credited with 67% consistency. Thus, we define innovation, in the context of a random condition, as consistency minus 33%.

Figure 3.5 shows mean innovation scores for the four conditions that occurred in both experiments, using only data from the comparable blocks of the two experiments as previously described. The plotted data was also analyzed by an ANOVA of 2 ages $\times$ 4 conditions, which confirmed that there was a significant difference in innovation between the ages ($F_{1,56} = 10.2, p < .01$), with children having innovation scores almost twice as high as adults (26% *vs.* 15%).

---

comparison, while the children had only seen half of them. Also, the children had experienced one test block preceding the first test block that is included. However, this was as close to comparable as it was possible to achieve given the different procedures.

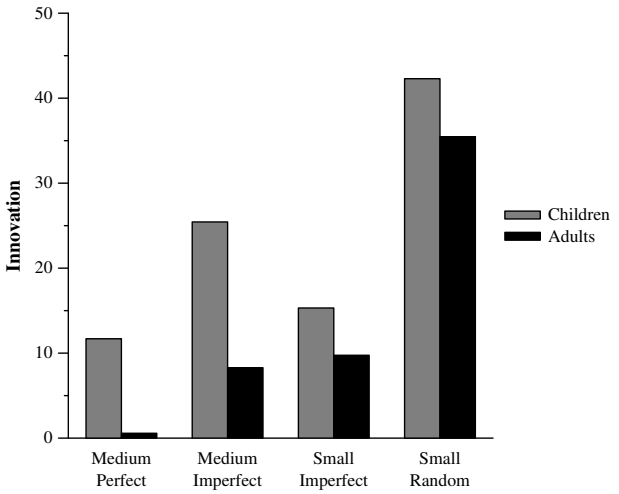

Figure 3.5: Comparison of innovation by adults and children. Data are from comparable conditions and blocks of Experiments 2 and 3 only. Children offer significantly more innovative responses in all conditions.

The difference between conditions was also significant ($F_{3,56} = 23.6, p < .001$); the Small Random condition led to the most innovation (46%), significantly higher than the next highest condition ($F_{1,56} = 30.0, p < .001$). Next was Medium Imperfect (17%), Small Imperfect (12%), and finally Medium Perfect (6%). These three were not reliably different pairwise, but Medium Imperfect was significantly larger than Medium Perfect ($F_{1,56} = 4.2, p < .05$). This confirms the intuitive idea that, for both adults and children, innovation of new patterns is is most likely when the patterns in the data are lacking or are difficult to find.

There was no interaction between the variables, which suggests that the bias of children toward innovative responding is not limited to particular kinds of data, but rather, it shows up in all of the combinations tested.

3.4 Discussion

One major finding of this analysis was that, despite the fact that children scored much lower than adults in terms of accuracy, the two age groups were nearly equal in terms of the consistency of their responses. The difference is made up by the much larger amount of innovation among the children (nearly twice as much overall).

The different conditions also revealed effects on consistency quite different from their effects on accuracy. The effect of data quality on accuracy was generally positive for both adults and children; both found the exception-free data easier to learn from. This same pattern holds for the effect of data quality on adults' consistency, but not for children. Children had nearly equal levels of consistent responding whether the data was perfect or random; the difference in accuracy was made up for by increased innovation. The only exception was in the Small Perfect condition, where accurate responding finally took over from invented patterns.

In general, for adults' performance, the use of the consistency measure does not add much to what we already had discovered. Except in the case of two or three subjects in the Small Random condition, there seems to be little tendency for adults to respond consistently according to patterns other than the ones suggested by the observation trials. If one were to paraphrase the rule that adults seem to be following, it

would be simply to match the observation data as closely as possible; probability-matching, with perhaps some bias towards exploratory or random responding to account for the tendency to undermatch. It appears that individual subjects have different responses to imperfect or random data, leading some to maximize and others to guess in an apparently random fashion.

Children, too, will follow the observation patterns if they find them. As would be expected due to their earlier stage of cognitive development, they are less successful in finding the patterns. However, they do show good learning when the target system is relatively simple, as is particularly evident in the Small Perfect system.

However, superimposed on this pattern-following behavior there is another response rule that children seem to follow just as strongly, which could be paraphrased as "Respond consistently." This rule applies not only in the random condition, but any time that the child has not (yet) figured out the accurate pattern. The made-up patterns seem to be followed with a consistency averaging around 70%, but ranging up to 100% consistency.

For example, consider again table 3.1. The child is responding to the observation-data's pattern only in the last block. In blocks 1 and 2, she responded according to innovated patterns with just about the same consistency. This illustrates another point: although accuracy tends upwards over blocks in many conditions, consistency appears to be stable over time; some subjects, like this one, are actually replacing innovated patterns with accurate ones, while maintaining a constant level of consistency.

4 Conclusions

This investigation began from curiosity about whether the mechanisms used for language acquisition and for learning other complex systems are the same or different. Linguistic and non-linguistic learning tasks are treated quite differently in the literature: while descriptions of language acquisition emphasize how children overlook exceptions and form rules even from inconsistent data, descriptions of *non*-linguistic tasks using probabilistic data usually highlight subjects' probability-matching behavior.

Various phenomena in language acquisition point to children as being the innovators and regularizers of language. Overregularization is a common example, showing that children are, for a time, willing to trust a generalization (past tenses are formed by adding "-ed") over specific conflicting data (adults use "went," not "goed"). Similarly, children exposed to inconsistent input from parents who are not native speakers can surpass that input, removing the irregularities (Singleton & Newport, 1993). And finally, creolization, the process by which children make a communication system into a language by creating and sharpening regularities, has been described as a more extreme example of this same general phenomenon: that children, but not adults, have a tendency to regularize linguistic systems that they encounter (Bickerton, 1984).

The most familiar result of probability-learning experiments, on the other hand, is choosing the alternatives just as frequently as they have been observed to be correct. However, such matching is not the universal outcome of probability-learning experiments. As was discussed in the introduction, matching appears to be only one of a number of possible techniques that learners have at their disposal: overmatching is common, slight differences in procedure can cause people to maximize rather than match, and in many cases matching behavior turns out, on more careful analysis, to actually be the result of the subject testing out hypotheses about patterns. Most interesting in terms of the present study is that children also behave differently than adults; certain ages (depending on the specific problem) are particularly likely to maximize, as well as more likely to *minimize* by following incorrect patterns (Weir, 1964; Bever, 1982).

Thus, although learning in the two domains has been described quite differently, the results do not seem incompatible when they are directly compared. Considering the large differences in behavior that can result from small changes in experimental procedure or system structure, which are well documented in the probability-learning literature, it is plausible to consider that the differences in learning seen between linguistic and non-linguistic tasks are due more to the structure of the problems that have been tested than to the difference in domain.

The systems taught in non-linguistic probability-learning experiments are usually extremely simple, often with only two or three possible responses and no context or stimuli that give information about the trial. Although rather sophisticated problems can certainly be devised even in this restricted paradigm (by making the probabilities vary over time, for example, or making the correct choice dependent on previous choices), the kind of complexity that is introduced is quite different from the kind of complexity that languages embody.

Maximizing:	Pick A_1 on (nearly) every trial.
Overmatching:	Pick A_1 with probability greater than p_1.
Matching:	Pick A_1 with probability p_1.
Undermatching:	Pick A_1 with probability less than p_1, but more often than the other responses.
Guessing:	Pick all responses approximately equally often.
Minimizing:	Pick A_1 less often than the other choices.

Table 4.1: Possible results of probability-matching experiments. Here A_1 stands for the most-often-correct response, and p_1 for the proportion of trials on which $A_1$is correct.

This study set out to test the apparent parallels between the two literatures, exploring what looked like commonalities between language and non-language tasks requiring learning from inconsistent data. The experiments follow probability-learning tasks in format, but use target systems that are constructed in ways analogous to linguistic structures. This new experimental paradigm worked well both with adults and with 7-year-old children, hitting a level of difficulty that was learnable by both groups, but challenging enough to show strong effects of the parameters that were varied.

The systems used were miniature analogues of morphology. The subjects learned to make appropriate connections between objects and actions, as language learners must find correspondences between words (or parts of words) and their meanings. In addition, the experimental systems were compositional, in that the mappings were best described at a fine level of detail (individual features of the objects and actions), but were presented as complete examples containing multiple pieces of information. Generalization trials were used to directly test whether subjects were sensitive to this fine level of detail; in order to generalize, subjects had to have analyzed the observation stimuli into their component parts, learned (or innovated) mappings at that level of detail, and then reassembled the resulting features in order to form complete responses.

If linguistic and non-linguistic learning are accomplished by the same mechanisms, and our system is close enough to linguistic systems in structure and complexity, then one would predict that some of the features characteristic of language-learning would also appear in the learning of this non-linguistic data set. Specifically, one might expect children to be more apt than adults to regularize imperfect patterns. If it is the complexity of language, in particular, that leads children to regularize, then the more complex versions of the experimental system might also be more clearly maximized.

Since the procedures used were designed to follow the general procedures of the probability-learning paradigm, let us first consider the results in those terms. In general, the results of a probability-learning experiment are described by comparing the proportion of trials on which a subject chooses a particular response to the proportion of trials on which that response was correct (see table 4.1).

Our hypotheses were framed in terms of the prevalence of maximizing, as opposed to matching, but the actual results turned out to be more varied. Matching and maximizing were both found, but to greater or lesser extents so were all of the other possibilities listed in table 4.1. The most common result was undermatching—responding with accuracies somewhere between chance and the data's consistency (quite the opposite of the prevalent tendency in probability-learning experiments for subjects to overmatch). There were also a few subjects who who maximized one dimension of the problem they were given, while apparently guessing on the others.

Adults' average performance was in some cases very close to the probability-matching level (particularly in the imperfect conditions of Experiment 2) , which makes it tempting to broadly summarize adults as matchers on this task. However, given the large individual differences, this may be an oversimplification.

It is possible that the differentiation of response probabilities is a general reaction to complex probability-learning tasks. In comparing these results to other probability-learning experiments, it must be kept in mind that the task in this experiment is considerably more intricate than the standard paradigm. Even the

simplest versions used here, the small systems, are comparable to running in three 3-choice probability-learning experiments simultaneously: there are three contexts (signaled by the three stimulus actions), in each of which the subject must learn the probabilities of each of the objects being the correct answer. In the medium and large systems, there are not only more contingencies to learn, but the correct mapping between dimensions must be figured out before the mappings between features can be considered.

Perhaps we should not be too surprised, therefore, that accuracies were for most subjects lower than matching level; even if probability-matching is the underlying response rule, there are many chances for failures of discrimination, learning, or memory to affect performance. A useful avenue for future research would be to try some systems intermediate in complexity between the standard probability-learning paradigm and the simplest systems used here, in order to find out just where and how the familiar probability-matching patterns begin to break down into results like those of the current study. In the absence of such studies, no thorough analysis of the data in terms of probability-learning theory is attempted here.

However, there are several descriptive generalizations that can be offered about adults' approach to this task: adult's responses appear to be probabilistic in nature, easily described (at the individual level, at least) within the vocabulary of probability-learning theory. Subjects' accuracy is directly related to the quality of their data, and consistency is nearly entirely due to accuracy. There is, however, great individual variation on just where in the spectrum from random guessing to maximizing they fall; this varies even within subjects on different parts of the system. Overall, the pattern appears to be that adults match or maximize parts of the system that they are confident of, and otherwise guess randomly.

These generalizations do not hold for children. In fact, there does not seem to be any satisfactory explanation of children's behavior using only the terminology of probability-learning theory. Rather, consistency and innovation are the unifying principles of children's approach to this task. While innovation falls under the general category of "minimizing," consistency crosscuts the distinctions; the same level of consistency can be found whether the subject is minimizing, maximizing, or anything in between. Thus, although the usual way of examining the results of probability-learning experiments is to look at accuracy scores, making sense of children's behavior required a different metric, which looks at the consistency of children's response patterns, regardless of the match between those patterns and the observation data.

A natural interpretation of the consistency results is easier to reach when one remembers that the systems were designed to be language-like in several respects, and considers the results in terms of the differences between adults and children in language acquisition. In linguistic tasks, there exists the possibility of learners acquiring something *different* than the target system, of enhancing or regularizing a pattern which changes the language. While we initially had not suspected that this might happen in a non-linguistic task, it became clear that this is what the children were doing with the systems we presented to them. Children in all conditions of the experiment seemed unsatisfied with random guessing; rather, they required that their responses follow some pattern consistently.

This level of consistent responding, which was around 60% for medium systems and 70-75% for small systems, was unaffected by the data quality or the accuracy of the children's responses; even the correlation between individual subjects' accuracy and consistency scores is small ($r = .71$, with a slope of .45; compare this to $r = .92$ for adults, with a slope of .76). This is impressive, given that accuracy is a direct constituent of consistency.

Children were able to have equal consistencies across the different levels of accuracy and data quality because of the phenomenon we have called *innovation*. Overall, children's responses seem to follow one of two strategies: if they can learn the system and respond accurately, they do so. But if not, rather than guessing, they innovate a response rule, and follow that on 60–75% of the trials. Compared to adults, they are much less efficient at extracting the actual rules (if any) from the data, but they also seem more willing to draw conclusions from data that is inconsistent or even totally random.

To return, then, to the question asked at the outset: does language acquisition recruit the same learning mechanisms as other learning? The evidence from these experiments does not support the claim in the way we had initially thought; we did not find clear maximizing by children where adults probability-match. Adults sometimes probability-matched and sometimes maximized, and children seldom did either one. Thus, in terms of our original hypotheses, the experiments proved inconclusive. If clear maximizing can be obtained from children in this kind of task, then it must be with a different sort of system.

However, the results in terms of consistency and innovation are more intriguing than even the predicted results would have been. Children, in a manner more subtle than was anticipated, demonstrated a determination to maintain a particular level of consistency in their use of these probabilistic systems. When unable to extract a pattern from the data they were given (or if there was none), they were quite willing to use a pattern created out of whole cloth.

These results are reminiscent of the innovative character of children's language acquisition, suggesting that there may indeed be similar mechanisms at work here. However, much work remains before these results will be more than suggestive. Further experimentation, designed particularly to look at this kind of innovative responding, is called for to answer the many questions that come to mind:

- Where, if anywhere, do children's "innovative" response rules come from? Are they picked at random, or is there something in the data, such as perhaps the first occurrence of a particular stimulus, that holds special salience for children and which is influencing their choice of patterns?
- At what age is this phenomenon most prevalent? 7-year-olds are already old enough that if they begin to learn a language they will do so differently than natives (Johnson & Newport, 1989; Krashen, Long & Scarcella, 1982). Thus, it would be particularly interesting to run a variant of this procedure with younger subjects. Indeed, the results of Bever (1982) and Weir (1964) suggest that maximizing is most prevalent at certain ages, but the structure or complexity of the task effects what ages those are.
- Why was matching not found in this experiment? It would be worth trying a similar procedure using simpler systems, to find out when and how probability-matching behavior breaks down into the highly variable behavior found among adults in the present study.
- A few adults appeared to be more similar to the children than to the other adults. Are these individual differences correlated with other differences in their learning styles, perhaps even in their language-learning ability? One of the puzzling results of critical period studies of second language learning is that adults show large variability: some adults look much more like childhood learners than do others (Johnson & Newport, 1989). Perhaps these individual differences in language learning may show up in laboratory paradigms as well, and are the resuls of individual differences in learning style more generally.
- What is it that causes adults and children to be different in this way? Why do children not seem to learn that the data *do not* follow any pattern?

These data suggest that children may innovate in the learning of any structured system, just as they are known to do in language acquisition. What, then, accounts for language's seeming uniqueness—why is it an exception to some "laws of learning"? Certainly imperfect data are found in every domain that children encounter; the world is not a perfectly consistent place. However, in learning about most types of data—arithmetic, say—any "innovative" pattern that children might use is simply wrong. We could speculate that what is unique about language is not any difference in the way children approach it, but a difference in the way language behaves when it is approached: adults find children's speech interesting and charming, even when it is quite ungrammatical, because of its consistent inventiveness; we make more of an effort to understand than to correct. Once in a while, a child's invention is so compelling that it is incorporated into the mainstream language—as if the language itself were too charmed to refuse.

In this way, languages adapt to the way children learn; generation by generation, the structure of the language comes more and more to resemble the biases of its learners, until the two are difficult to tell apart.

References

Baclawski, K., Rota, G.-C., & Billey, S. (1989). *An Introduction to the Theory of Probability* (2nd Preliminary ed.). MIT.

Bever, T. G. (1982). Regression in the service of development. In T. J. Bever (Ed.), *Regression in Mental Development: Basic Properties and Mechanisms* (pp. 153–88). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bickerton, D. (1984). The language bioprogram hypothesis. *The Behavioral and Brain Sciences*, *7*, 173–221.

Bogartz, R. S. (1965). Sequential dependencies in children's probability learning. *Journal of Experimental Psychology*, *70*(4), 365–70.

Bogartz, R. S. (1969). Short-term memory in binary prediction by children: Some stochastic information processing models. In G. H. Bower & J. T. Spence (Eds.), *Psychology of Learning and Motivation*, volume 3. NY: Academic Press.

Bower, G. H. & Hilgard, E. R. (1981). *Theories of Learning* (5th ed.). Century Psychology Series. Englewood Cliffs, NJ: Prentice-Hall.

Brookshire, K. H. (1976). Vertebrate learning: Evolutionary divergences. In R. B. Masterson, C. B. G. Campbell, M. E. Bitterman, & N. Hotton (Eds.), *Evolution of Brain and Behavior in Vertebrates* chapter 10, (pp. 191–216). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chomsky, N. (1975). *Reflections on Language*. NY: Random House.

Couvillon, P. A. & Bitterman, M. E. (1986). Performance of honeybees in reversal and ambiguous-cue problems: Tests of a choice model. *Animal Learning and Behavior*, *14*(3), 225–231.

Couvillon, P. A. & Bitterman, M. E. (1991). How honeybees make choices. Manuscript, University of Hawaii.

Estes, W. K. (1972). Research and theory on the learning of probabilities. *Journal of the American Statistical Association*, *67*(337), 81–102.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review*, *83*(1), 37–64.

Feldman, J. (1961). Simulation of behavior in the binary choice experiment. In *Proceedings of the Western Joint Computer Conference*, (pp. 133–44).

Friedman, M. P. et al. (1964). Two-choice behavior under extended training with shifting probabilities of reinforcement. In R. C. Atkinson (Ed.), *Studies in Mathematical Psychology* (pp. 250–316). Stanford University Press.

Gallistel, C. R. (1990). *The Organization of Learning*. Series on Learning, Development, and Conceptual Change. Cambridge, MA: MIT Press.

Gardner, R. A. (1957). Probability-learning with two and three choices. *American Journal of Psychology*, *70*, 174–85.

Gittins, J. C. (1989). *Multi-Armed Bandit Allocation Indices*. NY: Wiley.

Goldowsky, B. N. & Newport, E. L. (1993). Modelling the effects of processing limitations on the acquisition of morphology: The Less Is More hypothesis. In Clark, E. V. (Ed.), *Proceedings of the Twenty-Fourth Annual Child Language Research Forum*, volume 24, Stanford, CA. Stanford University, CSLI.

Grant, D. A., Hake, H. W., & Hornseth, J. P. (1951). Acquisition and extinction of a verbal conditioned response with differing percentages of reinforcement. *Journal of Experimental Psychology*, *42*(1), 1–5.

Harper, D. G. C. (1982). Competitive foraging in mallards: "Ideal free" ducks. *Animal Behavior*, *30*, 575–84.

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior*, *13*(2), 243–266.

Herrnstein, R. J. & Vaughan, Jr., W. (1980). Melioration and behavioral allocation. In J. E. R. Staddon (Ed.), *Limits To Action: The Allocation of Individual Behavior* chapter 5, (pp. 143–176). NY: Adademic.

Humphreys, L. G. (1939). Acquisition and extinction of verbal expectations in a situation analogous to conditioning. *Journal of Experimental Psychology*, *25*, 294–301.

Johnson, J. S. & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60–99.

Kay, P. & Sankoff, G. (1974). A language universals approach to pidgins and creoles. In D. DeCamp & I. F. Hancock (Eds.), *Pidgins and Creoles: Current Trends and Prospects*. Washington DC: Georgetown.

Krashen, S. D., Long, M. H., & Scarcella, R. C. (1982). Age, rate, and eventual attainment in second language acquisition. In S. D. Krashen, R. C. Scarcella, & M. H. Long (Eds.), *Child-Adult Differences in Second Language Acquisition*. Rowley, MA: Newbury House.

Kyburg, H. E. (1990a). Probabilistic inference and non-monotonic inference. In R. D. Shachter, T. S. Levitt, L. N. Kanal, & J. F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 4* (pp. 319–326). North-Holland: Elsevier Science Publishers B.V.

Kyburg, H. E. (1990b). Probabilistic inference and probabilisitc reasoning. *Philosophical Topics*, *18*(2), 107–116.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57*(4), 1–165.

Newport, E. L. (1981). Constraints on structure: Evidence from American Sign Language and language learning. In *Aspects of the Development of Competence.*, volume 14 of *Minnesota Symposia on Child Psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Newport, E. L. (1982). Task specificity in language learning? Evidence from speech perception and American Sign Language. In E. Wanner & L. R. Gleitman (Eds.), *Language Acquisition: The State of the Art*. NY: Cambridge Univ.

Newport, E. L. (1984). Constraints on learning: Studies of the acquisition of American Sign Language. *Papers and Reports on Child Language Development*, *23*, 1–22.

Newport, E. L. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of American Sign Language. *Language Sciences*, *10*, 147–172.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11–28.

Newport, E. L. (1991). Contrasting conceptions of the critical period for language. In S. Carey & R. Gelman (Eds.), *The Epigenesis of Mind: Essays on Biology and Cognition*. Hillsdale: Erlbaum.

Reber, A. S. & Millward, R. B. (1968). Event observation in probability learning. *Journal of Experimental Psychology*, *77*(2), 317–327.

Reber, A. S. & Millward, R. R. (1971). Event tracking in probability learning. *American Journal of Psychology*, *84*, 85–99.

Rescorla, R. & Wagner, A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory*. NY: Appleton-Century-Crofts.

Restle, F. (1967). Grammatical analysis of the prediction of binary events. *Journal of Verbal Learning and Verbal Behavior*, *6*, 17–25.

Ross, B. M. & Levy, N. (1958). Patterned predictions of chance events by children and adults. *Psychological Reports*, *4*, 87–124.

Sankoff, G. (1979). The genesis of a language. In K. C. Hill (Ed.), *The Genesis of Language*. Ann Arbor: Karoma Publishers.

Sankoff, G. & Laberge, S. (1973). On the acquisition of native speakers by a language. *Kivung*, *6*, 32–47. Reprinted in DeCamp and Hancock.

Schwartz, B. & Reisberg, D. (1991). *Learning and Memory*. NY: W. W. Norton.

Singleton, J. L. (1989). *Restructuring of Language from Impoverished Input: Evidence for Linguistic Compensation*. PhD thesis, University of Illinois at Urbana-Champaign.

Singleton, J. L. & Newport, E. L. (1993). When learners surpass their models: The acquisition of American Sign Language from impoverished input. Manuscript under review.

Staddon, J. E. R. (1980). Optimality analyses of operant behavior and their relation to optimal foraging. In J. E. R. Staddon (Ed.), *Limits To Action: The Allocation of Individual Behavior* chapter 4, (pp. 101–141). NY: Adademic.

Suppes, P. & Atkinson, R. C. (1960). *Markov Learning Models for Multiperson Interactions*. Stanford University Press.

Vitz, P. C. & Todd, T. C. (1967). A model for simple repeating binary patterns. *Journal of Experimental Psychology*, *75*(1), 108–17.

Weir, M. W. (1964). Developmental changes in problem solving strategies. *Psychological Review*, *71*, 473–490.