The Dissertation Committee for Elizabeth (Lisa) Christine Kaczmarczyk
certifies that this is the approved version of the following dissertation:

# The Acquisition of Intellectual Expertise: A Computational and Empirical Theory

Committee:

---

Risto Miikkulainen, Supervisor

---

Andrew Bernat

---

Anthony Petrosino

---

Raymond Mooney

---

Lowell Bethel

---

Bradley Love

# The Acquisition of Intellectual Expertise: A Computational and Empirical Theory

by

**Elizabeth (Lisa) Christine Kaczmarczyk, B.A.;M.S.;M.S.**

**Dissertation**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Doctor of Philosophy**

**The University of Texas at Austin**

August 2005

[0] *Technical Report, The University of Texas at Austin Department of Computer Sciences UT-AI-TR-05-321*,2005

To Dandelion Kaczmarczyk, who always reminded me about the most important things in life.

# Acknowledgments

There are so many people who supported, encouraged and mentored me while I worked on this dissertation. Most important, I would like to thank my advisor, Risto Miikkulainen. First, for supervising this interdisciplinary research. Second, for teaching me so much about conducting rigorous research and expressing my results with confidence. Third, for being a nice person. I would also like to thank the other members of my committee. Andrew Bernat, for his excellent advice on many occasions; Anthony Petrosino, for directing me to important resources on cognition and learning; Raymond Mooney, for his perspective on machine learning; Bradley Love for his perspective from cognitive psychology; Lowell Bethel, for his encouragement, especially during my early years at UT.

Many other people supported me and my work at critical junctures. I want to especially thank Marilla Svinicki for her support during my comprehensive exams, and when I was developing my human subject study. Also, Jim Bednar, for his technical advice on numerous occasions, and Elaine Rich for her understanding of the importance of teaching and learning. Many of the staff in the Computer Sciences department are terrific people who made my life easier in countless ways. My parents, Patricia Harris and Alexander Kaczmarczyk, provided crucial financial assistance. David Alexander listened, provided perspective, and challenged me; This dissertation would not be the same without our conversations. Finally, I must thank my wonderful friends, who provided emotional, logistical and physical support, and cheerfulness! There are too many of you to list individually, but each of you has a special place in my memory.

Elizabeth (Lisa) Christine Kaczmarczyk

*The University of Texas at Austin*

*August 2005*

# The Acquisition of Intellectual Expertise: A Computational and Empirical Theory

Publication No. ______________________

Elizabeth (Lisa) Christine Kaczmarczyk, Ph.D.
The University of Texas at Austin, 2005

Supervisor: Risto Miikkulainen

In order to develop intellectual expertise, a novice learner has to acquire cognitive abilities seen in experts: They need to be able to categorize problems before solving them, and be meta-cognitive about their learning, so that they can select the best problem-solving strategies. The process by which learners acquire these abilities is not well understood. The goal of this dissertation is to understand how instructional delivery methods can help learners acquire the cognitive abilities necessary to become experts. The study was motivated by initial structured interviews with mathematics faculty, which led to the formulation of three hypotheses: (1) Traditional sequential delivery methods inhibit learning and retention; (2) Integrated delivery methods increase learning and retention; (3) Incrementally increasing the complexity of the material will lead to the best performance. An artificial neural network was then used to test these hypotheses computationally. The network confirmed the hypotheses, demonstrating that an Incremental delivery leads to better learning than Drill and Test learning or Fully Integrated learning. These computational conclusions led to the prediction that an Incremental Learning delivery method will encourage meta-cognitive abilities necessary to achieve expertise. This prediction was tested experimentally on human subjects. Qualitative and quantitative data from the human study verified that (1) Incremental learners develop the most effective study and test taking strategies; (2) Incremental learners have the best conceptual development; and (3) Incremental learners have the most positive reactions to learning. I hope that these results will benefit society, because by changing the way we educate students, more learners can pursue advanced study, and use their expert, creative insights to address society's most challenging problems.

# Contents

**Acknowledgments** **v**

**Abstract** **vii**

**List of Tables** **xiii**

**List of Figures** **xiv**

**Chapter 1 Introduction** **1**

1.1 How is Intellectual Expertise Acquired? . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.2 Acquiring Calculus Expertise . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.3 Approach . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

1.4 Outline of the Dissertation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

**Chapter 2 Background** **9**

2.1 Learning and Expertise as Cognitive Performance . . . . . . . . . . . . . . . . . . . . . 9

2.2 Learning and Expertise as a Developmental Process . . . . . . . . . . . . . . . . . . . 12

2.2.1 Connectionist Models of Learning and Cognition . . . . . . . . . . . . . . . . . 13

2.2.2 Meta-cognition as reflected in Strategy Development . . . . . . . . . . . . . . 15

2.3 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

**Chapter 3 Research Questions** **18**

3.1 Motivation . . . . . . . . . . 19
3.2 Interviews with Experts . . . . . . . . . . 21
3.2.1 Overview of the Grounded Theory Method . . . . . . . . . . 21
3.2.2 Participants and Procedure . . . . . . . . . . 23
3.2.3 Results . . . . . . . . . . 23
3.2.4 An Emerging Theory of Instruction For Expertise . . . . . . . . . . 25
3.3 Development of Approach . . . . . . . . . . 26
3.3.1 Three Hypotheses to be Tested . . . . . . . . . . 26
3.3.2 Why a Connectionist Model? . . . . . . . . . . 28
3.4 Conclusions . . . . . . . . . . 28

**Chapter 4 Testing the Hypotheses Computationally** **30**
4.1 The Artificial Neural Network Model . . . . . . . . . . 30
4.1.1 Architecture and Data . . . . . . . . . . 30
4.1.2 Description of the Input Vector . . . . . . . . . . 32
4.1.3 Experimental Design . . . . . . . . . . 36
4.2 Computational Experiments . . . . . . . . . . 38
4.2.1 Drill and Test Learning Simulations . . . . . . . . . . 38
4.2.2 Fully Integrated Learning Simulations . . . . . . . . . . 39
4.2.3 Incremental Learning Simulations . . . . . . . . . . 39
4.3 Discussion . . . . . . . . . . 44
4.4 Conclusions and a Prediction . . . . . . . . . . 47

**Chapter 5 Testing the Computational Prediction with Human Subjects** **48**
5.1 Motivation . . . . . . . . . . 48
5.2 Methodology . . . . . . . . . . 50
5.2.1 Overview of the Meaning Categorization Method . . . . . . . . . . 50

5.2.2 Subjects and Materials . . . 54
5.2.3 Experimental Procedure . . . 55
5.2.4 Experimental Analysis . . . 56
5.3 Qualitative Analysis . . . 60
5.3.1 Drill and Test . . . 60
5.3.2 Fully Integrated . . . 61
5.3.3 Incremental Learning . . . 62
5.4 Quantitative Analysis . . . 64
5.4.1 Learning . . . 64
5.4.2 Performance . . . 66
5.5 Discussion . . . 67
5.5.1 Integrating Qualitative and Quantitative Results . . . 67
5.5.2 Affect and Motivation . . . 68
5.5.3 Open Questions and Future Directions . . . 69
5.6 Conclusions . . . 70

**Chapter 6 Discussion and Future Work** **71**

6.1 Relationship to Behaviorism and Constructivism . . . 71
6.2 Conceptual Development: Hidden Layer Analysis . . . 73
6.3 Conceptual Development: Error Analysis . . . 75
6.4 Prior Knowledge Through Analogy . . . 76
6.5 Affect and Motivation . . . 77
6.6 Computational Experiments in Science Domains . . . 78
6.7 Extensions to the ANN Model . . . 79
6.7.1 Generating New Neural Connections . . . 79
6.7.2 Linking Integration and Differentiation . . . 79
6.7.3 Solving the Integration Problems . . . 81

6.8 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 82

**Chapter 7 Conclusion** **83**

**Appendix A Expert Interview Questions** **86**

**Appendix B Human Subject Experiment Interview Questions** **88**

**Bibliography** **90**

**Vita** **101**

# List of Tables

4.1 Representation of Constants and Variables . . . . . . . . . . . . . . . . . . . . . . . . 33

4.2 Representation of Unary Operators . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

4.3 Representation of Binary Operators . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35

5.1 Code List - Interview Analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 57

5.2 Differences in Strategy Development, Conceptual Development, and Affective Reactions in Drill and Test, Fully Integrated, and Incremental Learning. . . . . . . . . . . . 65

5.3 Human Performance. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 66

# List of Figures

4.1 The Computational Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31

4.2 The Integration Problems . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37

4.3 Drill and Test Learning Performance . . . . . . . . . . . . . . . . . . . . . . . 40

4.4 Drill and Test Learning by Concept . . . . . . . . . . . . . . . . . . . . . . . 41

4.5 Fully Integrated Learning Performance . . . . . . . . . . . . . . . . . . . . . . 42

4.6 Fully Integrated Learning by Concept . . . . . . . . . . . . . . . . . . . . . . 43

4.7 Incremental Learning Performance . . . . . . . . . . . . . . . . . . . . . . . . 45

4.8 Incremental Learning by Concept . . . . . . . . . . . . . . . . . . . . . . . . . 46

# Chapter 1

# Introduction

Experts can solve complex problems quickly and intuitively. For example, consider the following calculus integration problems:

$$\int 2x\sqrt{1+x^2}\mathrm{d}x$$
$$\int_0^{\pi/2}(x+3\cos(x))\mathrm{d}x$$
$$\int x^2\cos(3x)\mathrm{d}x$$
$$\int x^3\cos(x^4+2)\mathrm{d}x$$

A mathematician can immediately tell that the calculus integration problem $\int_0^{\pi/2}(x+3\cos(x))\mathrm{d}x$ can be solved by Simple integration, that $\int 2x\sqrt{1+x^2}\mathrm{d}x$ and $\int x^3\cos(x^4+2)\mathrm{d}x$ can be solved by Usubstitution, and that $\int x^2\cos(3x)\mathrm{d}x$ requires Integration by Parts.

For non-experts, such decisions are very hard to make. Yet such expertise is not innate, nor a general ability, but is learned (Ericsson and Smith 1991). The goal of this dissertation is to understand how instruction can help learners acquire the cognitive abilities necessary to become an intellectual expert. The term "intellectual expertise" refers to expertise that is primarily cognitive, rather than physical or biomechanical (e.g. typing or gymnastics). Intellectual expertise is complex, but can be measured with performance criteria and by monitoring meta-cognitive development (how learners monitor and regulate their own thinking, Flavell (1976). In order to establish a link between instructional method and cognition, the learning process has to be modeled and verified using an interdisciplinary approach. In this chapter, the problem of learning for expertise in general and calculus learning in particular are introduced. The advantages of using an interdisciplinary approach to study human learning will be outlined. Based on these facts, I will motivate the approach and conclude with an overview of the dissertation.

## 1.1 How is Intellectual Expertise Acquired?

An intellectual expert has achieved a level of performance in a problem domain such that she or he can rapidly grasp subtleties of complex problems, and produce high-quality solutions (Dörner and Schölkopf 1991). As a result of studies on expertise, we know a lot about what expert behavior is like. For example, expert chess players can evaluate board positions more accurately and select better moves than novices (literature reviewed by Charness (1991)). Similarly, physics experts represent problems in abstract terms that lead to good solutions, whereas novices use naive visual cues that lead to poor solutions (literature reviewed by Anzai (1991)).

However, we do not know much about how people become experts. It is not easy to create expertise, whether human or computational. The learning process is complex and human studies are difficult. Understanding this process has proven elusive for educators, psychologists and students alike.

It is particularly difficult to study expertise learning in domains that are ill-structured, such as advanced mathematics. In such domains, the concepts and cases are complex, and not all cases of a concept share the same surface-level structure (Duffy et al. 1993). In addition, there is often more than one way to solve a problem, however some methods are very inefficient. Furthermore, problems in ill-structured domains are incomplete, missing a clear definition of the goal state, the initial state, the operators, or the constraints (Reitman 1964). As a result, it is difficult to design learning tasks that are both generalizable and realistic. Expertise takes many years to develop (Chase and Simon 1973; Ericsson et al. 1993). Longitudinal studies, even when feasible to conduct, can not control for the many influences that are part of daily living.

As a result, much of what we know about acquiring expertise is based upon retrospective analyses of the history of current experts, comparisons of current novices to experts, or developmental studies with infants and children. Such limited knowledge is a problem when designing methods for instructing people to become experts. To be most effective, instructional decisions need to be based upon verifiable principles tested using the domain where they will be applied, in this case calculus. Therefore, an expertise learning problem needs to be selected from the calculus domain.

## 1.2 Acquiring Calculus Expertise

Calculus is a good domain to study expertise learning for three reasons. First, in order to become calculus experts, students need to understand complex concepts and intuitively select the most efficient methods to solve problems. In high-pressure situations such as taking exams, or working on commercial projects, there is no time for trial and error problem-solving. Second, calculus is a topic students must master to be successful in many of the natural sciences and engineering. Any research results may therefore apply to future experts in many fields that depend upon high mathematics performance. Third, calculus, at its most fundamental level, is based upon abstract cognitive concepts (e.g. limit, instantaneous rate of change). As a result, understanding how people best learn calculus requires understanding the mind. Bruner and Kenney (1965) even suggested that learning mathematics may be viewed as a microcosm of all intellectual development.

Calculus is an ill-structured domain, which makes it difficult for both instructors and students. For example, all four of the problems shown at the beginning of this chapter *can* be solved using Integration by Parts. However, Integration by Parts is the most complex of the three solution methods and is not the preferred choice when a simpler method will work. No one has produced an algorithm that an instructor can give a novice that will help her identify the most effective solution method.

When well-intentioned attempts to simplify complex material (Gagné 1968) result in novices being taught superficial concepts, these learners cannot master complex concepts and transfer knowledge to new problem situations (Duffy et al. 1993). The difficulty does not appear until the learner encounters advanced material. At that point, he or she discovers that problems nominally of the same type in fact have flexible category boundaries, and it is not clear how to solve them (Duffy et al. 1993) .

For example, consider again the four calculus integration problems used in the beginning of this chapter. In deciding what solution methods to apply, a calculus novice might rely on surface level features of the problems (Norman and Prichard 1994; Chi et al. 1981) and reason as follows:

*Three of the problems contain cos(x). Two of those three problems begin with x raised to a power. There is a good chance that those two problems should be solved using the same solution strategy. On the other hand, two of the cos problems are longer and contain the addition operator. Perhaps they should be solved using the same solution strategy.*

This line of thinking leads to a wrong conclusion in both cases. According to the calculus textbooks from which these problems were taken (Lang 1986; Silverman 1985; Stewart 1995) only the first and last problems

$$\int 2x\sqrt{1+x^2}\mathrm{d}x$$

$$\int x^3 \cos(x^4+2)\mathrm{d}x$$

share the same best solution strategy (integration using Usubstitution). The other two problems are best solved using Simple integration, and Integration by Parts. The surface level features can be deceiving.

Examples of faulty mathematical reasoning similar to that just described, have been demonstrated empirically. Norman and Prichard (1994) showed that many learners cannot interpret the structure of a problem beyond surface-level symbols. They suggested that novices have inaccurate intuitions about problems, which lead them to attempt incorrect solution strategies. Because they cannot see beyond high-level features, they cannot develop correct intuitions. Conversely, successful problem solvers can recognize (but not always describe) underlying structural similarities and fundamental principles, and use them to correctly categorize mathematics problems. (Schoenfeld and Herrmann 1982; Silver 1979). These categories are often grouped based upon solution strategies that the experts use to calculate the answer (Owen and Sweller 1989).

Fortunately, instructors can help students develop the expert cognitive abilities needed to identify correct solution strategies (Ericsson and Smith 1991). A major open question for expertise learning is what instructional methods will support the development of cognitive abilities necessary for expertise. There is experimental evidence that some teaching methods actually make learning calculus harder, and lead to the kind of problem demonstrated in the above example with four integration problems. For example, Selden et al. (1994) concluded that providing students with isolated, trivial problems, the norm in many classrooms, inhibits the students from learning to generalize calculus problem solving skills. We need to understand more about how instructional methods can help, rather than hinder, learning for expertise. The next section will discuss how this dissertation addresses this problem.

## 1.3 Approach

A primary goal of the research reported in this dissertation is to understand the process by which humans become intellectual experts, and in particular, how educational delivery methods help students develop the meta-cognitive abilities necessary for expertise. In order to accomplish this task, an interdisciplinary research approach is needed, taking advantage of the strengths of established investigative and experimental methods from computer science, psychology and education. The strengths of these methodologies will be reviewed in order to understand the scope of the dissertation.

Research on human cognition and learning takes place within many disciplines and at many levels, ranging from neurobiological analysis (Packard et al. 1996) to field observations (Schofield 1995) . Traditionally, cognitive psychology research has focused on tightly controlled and isolated phenomena. The strength of this approach is that tight control eliminates much irrelevant data. Results can be analyzed statistically and generalized. The primary limitation to this approach is that it is sometimes difficult to apply the results to situations that are not controlled, such as human classroom learning.

Many recent educational studies take a different approach, borrowing qualitative research methods from the social sciences (Barker et al. 2002; Nielson et al. 1997; Marton and Booth 1997; Erickson 1986). Such methods traditionally focus in great detail upon a small number of human subjects. The strength of this approach is that the results provide a complex, detailed and well-rounded description of phenomena. The primary limitation is that sometimes it is not possible to generalize the results beyond the small study population.

A third experimental approach for studying human learning is to conduct a computational study. Such studies develop a model simulation of a phenomenon. This approach has several strengths. First, computational models can provide insight not easily obtainable through conventional studies on humans, because many hypotheses can be tested on the same set of data (representing human learners). Second, successful simulations make a strong argument for human studies that otherwise might lack political or financial support. Third, simulations make it easier to study poorly understood phenomena, such as risky hypotheses, without risk to human subjects. Fourth, simulations can be used to compress time, allowing the researcher to perform longitudinal investigations that would otherwise be impractical or even impossible to do. The primary limitation to computational simulations of human learning is that it is hard to guarantee that they match humans and therefore the results should still be verified with human subjects.

The best way to study expertise learning is to systematically combine statistical, qualitative and computational studies. The advantages of each approach provide a unique perspective on human learning, and the limitations of each method are minimized. Therefore, this dissertation will present a series of computational and human subject studies of expertise learning. At their conclusion, the results provide multi-disciplinary support for a theory of expertise learning.

The first results reported in this dissertation are based on an interview study with University of Texas at Austin (UT) Mathematics faculty and teaching assistants (TAs). These interviews provide insight into the complex relationship between experts' understanding of their own problem solving process and the teaching strategies they chose. These insights are consistent with the psychological literature on expert vs. novice behavior. Novice learners (in this case UT students) are often unable to select the correct integration solution strategy. This fundamental problem arises before they even have a chance to exhibit computational difficulties and prevents many from reaching timely, correct solutions. Conversely, experts (in this case faculty and TAs) claimed the ability to “just see” the correct strategy, yet were unable to articulate how they knew.

The interview data was used to formulate three hypotheses which were then tested with the computational model. The second set of results reported in this dissertation come from a series of experiments with this model, focusing on how three delivery methods, Drill and Test (DT), Fully Integrated (FI), and Incremental Learning (IL) affect expertise learning. The model is an artificial neural network (ANN). An ANN was selected because connectionist models have provided insight into related areas of cognition such as memory (McClelland et al. 1995), mathematics learning (Cottrell and Tsung 1989) and strategy development (Bray et al. 1997). The model was validated by showing that Drill and Test and Fully Integrated results match well-established data from educational studies about human learning. The model was then used to make predictions about improved human performance using an Incremental Learning delivery method. Finally the model was used to make the prediction that an Incremental Learning delivery method will encourage meta-cognitive abilities seen in experts.

The third set of results conclude this dissertation by testing the computational prediction that an Incremental Learning delivery method will encourage meta-cognitive abilities seen in experts. The study evaluates the effect of the three delivery methods on human performance, strategy development and conceptual development. The experiment consisted of two parts: a formal laboratory study followed by a structured interview. The results of the human subject study were analyzed using quantitative and qualitative measures, and they support the prediction that an Incremental Learning delivery method leads to the best expertise learning.

Taken together, the results provide an interdisciplinary theory that an Incremental Learning delivery method is the best way to instruct for expertise learning. By using Incremental Learning in the classroom we can improve education. I hope that more students will excel in their studies and become experts in their chosen fields; Society will benefit from their insightful understanding of complex problems.

## 1.4 Outline of the Dissertation

This dissertation is organized into three parts: Introduction and Background (**Chapters 1** and **2**), Methodologies and Results (**Chapters 3** through **5**), Discussion, Future Work and Conclusion (**Chapters 6** and **7**).

**Chapter 2** will review the prior work on human and computational studies of expertise. The shift from an early focus on expert performance, to a more recent focus on the developmental process will be discussed.

**Chapter 3** will present results of an exploratory interview study with mathematics experts. These results motivated three experimental hypotheses about how instructional method could affect teaching for expertise. The hypotheses are: (1) Traditional sequential delivery methods inhibit learning and retention; (2) Integrated delivery methods increase learning and retention; (3) Incrementally increasing the complexity of the material will lead to the best performance.

**Chapter 4** describes the artificial neural network model in detail. The results of three sets of computational experiments will be presented, verifying the three hypotheses computationally. A prediction is then derived, that an Incremental Learning delivery method will encourage metacognitive abilities seen in experts.

**Chapter 5** discusses the results of a human subject experiment that tested the prediction.

**Chapter 6** discusses the most promising future directions in computational and human subject research of learning for expertise. These studies include investigation of conceptual development in the hidden layers of the ANN, use of analogy by human learners, and extensions to the architecture of the model so that additional psychological phenomena can be modeled.

**Chapter 7** summarizes and evaluates the contribution of the dissertation.

# Chapter 2

# Background

This chapter will discuss what we do and do not know about expertise and expertise learning. Research about intellectual expertise developed as a subfield of studies about learning and problem-solving. Early human and computational studies of expertise focused on analyzing and modeling the cognitive properties of experts. More recent studies have shifted the focus on expertise learning as a developmental process. Meta-cognition is now recognized as an important expert ability, and human studies can measure it by monitoring strategy development. Connectionist models, inspired by the structure of the human brain, have provided insight into many areas of human learning, including mathematics learning. As a result of computational and human studies, we know a lot about the differences between novices and experts. However, we still do not know enough about how instruction can help learners become experts.

## 2.1 Learning and Expertise as Cognitive Performance

Early studies of expertise framed problem-solving as a heuristic search problem (Newell et al. 1958; Newell and Simon 1972). These studies proposed that human experts had superior search abilities, which meant that they could apply their skills across many domains. As a result of this orientation, much expert problem-solving research focused on general reasoning and decision-making processes, including the early artificial intelligence (AI) programs. For example, the Logic Theorist (Newell et al. 1958) started with a set of axioms and applied a set of inference rules to them systematically until the theory was proved.

Another highly influential early AI program was the General Problem Solver (GPS); Newell and Simon (1963). The GPS established a complex theoretical framework of human cognition, making it one of the earliest computational models of cognitive processing. Like the Logic Theorist, GPS proved theorems by applying rules. GPS's primary innovation was to introduce the concept of production rules, which represent domain knowledge as a set of if-then clauses. These clauses were applied to a problem until the theorem was proved. The results of these simulations were compared to the results of human behavior on the same task.

From this time forward, it became increasingly popular to place the human mind and the computer in the same family of physical symbol systems. For example, Simon (1969) described memory and language as list structures operated upon by a serial processor. This conceptualization was highly influential in both the cognitive psychology and computer science research communities.

Although the early AI simulations contributed greatly to our understanding of logical reasoning, later expertise studies failed to confirm the theory that expertise is a generalized cognitive ability. Instead, seminal studies by DeGroot (1966), Chase and Simon (1973), and Chi et al. (1981) showed that expertise was domain-dependent. Experts often could not transfer their abilities from one domain to another, partly because they relied upon vast amounts of domain knowledge, acquired over many years (Ericsson et al. 1993; Ericsson and Charness 1994).

As a result of this discovery, expertise research began to focus on the role of memory. The influential theory of "chunking" grew out of this focus (Rosenbloom and Newell 1986). Simply put, chunking is a way of mentally organizing large amounts of knowledge, by dividing it into more manageable "chunks", that can be efficiently organized and referenced. Experts are posited to be better at performing chunking than novices, which leads to their superior ability to acquire, retain and use data. The theory of chunking is now a widely accepted description of expert ability.

Early AI programs that relied on production systems evolved into expert systems. Such systems formalize the decision-making process of a human domain expert by incorporating large amounts of expert knowledge. Examples typical of these systems can be found in medicine (Warren et al. 1993), law (Woodin 2001), engineering (Reed 2000), and reading comprehension (Dyer 1983). The input was typically a set of data describing a complex problem, and the goal was to determine the best solution by classifying the problem correctly. This process contained information about which specific sequences of actions will lead to desired solutions. To accomplish this task, the programs stored a vast amount of domain-specific data, and applied rules one at a time until a solution was reached. A successful expert system matched or exceeded the performance of a human expert on the same problem. Some of these computational systems evolved into grand theories of human cognition; two classic examples are SOAR (Laird et al. 1987), and the ACT theory (Anderson 1983).

The early human and computational expertise studies led to a relatively good understanding of what experts do. First, experts can see patterns in the domain better than non-experts. This ability allows them to think critically and solve problems in the domain (Chi et al. 1981; Murphy and Wright 1984). Second, we know that experts and novices categorize problems differently, and this categorization takes place before the expert attempts to solve the problem (Chi et al. 1981). Third, experts can correctly categorize problems without solving them (Hinsley et al. 1977; Robinson and Hayes 1978). Fourth, we know that there is a relationship between how experts decide to categorize problems and the solution strategies they will use to solve them. Specifically, the cognitive schema that experts develop to help them categorize problems are strongly influenced by the solution methods they prefer to use (Ross 1996, 1997). Finally, experts solve routine problems not by intensely calculating but rather by recognizing a type of problem (categorizing) and then using the stored knowledge about how to solve problems of that type (multiple references by Ross and Spalding (1991)).

However, expertise studies that focus on what an expert knows, rather than the process by which she or he attained the expertise (Duffy et al. 1993) are limited in some respects. First, early human expertise studies were often conducted in strictly controlled laboratory settings, in order to isolate phenomena and provide generalizable results. A side effect of this methodology is that the only domains studied were those where it was possible to clearly define tasks that comprise the expertise (such as chess and physics problem-solving). Many areas of expertise are not so easy to quantify, such as recognizing abstract mathematical concepts. Controlled laboratory studies also generally emphasize skill acquisition, not handling novel situations. This emphasis is a problem because experts are known to be very good at solving novel problems in their domain (Dörner and Schölkopf 1991). Therefore, it is not clear how far the results of the controlled studies can be extended. There is little empirical data to support the assumption that basic mental functions studied with traditional laboratory tasks are the same as those underlying performance in complex everyday life situations (Ericsson 2003).

Second, when expertise studies evaluate people who are already experts, analysis of the learning process becomes secondary. Understandably, people have sometimes made assumptions about how expertise learning occurs, but these assumptions may not be valid. We need to understand how expertise is learned, so that we can improve instructional methods. Therefore, the learning process itself needs to be the focus of rigorous study.

The AI expert systems are limited in the same way. For example, although expert systems are domain experts, they focus on reaching a goal state, and are less concerned with the learning process. If their purpose is, for example, to provide training or diagnosis for a commercial organization, this limitation is fine. However, if the goal is to better understand human expert learning, then the limitations are a problem, because they do not model learning as it occurs in humans. In particular, because expert systems depend upon logical rules, they do not perform well when solution paths are unclear, or when decisions need to be made based on statistical correlations. As these limitations have become clear, studies of human and computational problem-solving have significantly changed. Recent studies have shifted away from modeling expertise as a fixed cognitive state, and now focus on modeling cognitive developmental processes seen in humans.

## 2.2 Learning and Expertise as a Developmental Process

Studies of learning and expertise in the last two decades have branched out in many directions, such as conceptual development (Brewer 1993), creativity (Schunn and Anderson 1999; Bonnardel 1999), prior knowledge (Greene 1993), and motivation (Deci et al. 1991; Flowers et al. 1991). The common thread connecting these studies is the understanding that expertise learning is a developmental process and that expert decision-making often does not follow rules. In this dissertation I will focus on expertise learning assessed by performance and meta-cognition. A focus on performance is reasonable, because it is a well-established way to measure progress. A focus on meta-cognition is justified because meta-cognition is arguably the most important cognitive attribute of expert learners. As a reminder, meta-cognition is defined as self-appraisal and regulation of cognition by the learner (Flavell 1976). Studying meta-cognition will produce insight about expertise learning that will be useful to cognition researchers in many disciplines. Other areas for studying expertise learning computationally and empirically will be discussed in the Discussion and Future Work chapter.

This section will provide the background for understanding how the research presented in this dissertation builds upon recent computational and psychological advances in expertise learning. First I will discuss how connectionist AI models of human learning have framed problem-solving as a developmental process. In particular, two connectionist systems that simulated learning arithmetic will be described, to demonstrate the potential that artificial neural networks (ANNs) have in providing insight into expertise learning in mathematics. This background material will provide the context for understanding the computational model presented in this dissertation. Then, to provide the context for understanding the human subject experiment presented in this dissertation, I will explain why a significant amount of recent human subject research in expertise learning has focused on meta-cognition. Strategy development is a good way to monitor meta-cognition, and has been used to study expertise learning. Finally, this section discusses the limitations to previous studies of expertise learning, and why we need to conduct studies that focus on how instruction can help improve expertise learning.

### 2.2.1 Connectionist Models of Learning and Cognition

Recent advances in neuroscience have spurred increased interest in the development of parallel computational models of human learning. The brain has a parallel and highly interconnected structure; during learning neurons are physically rewired, and connection strengths change (Bransford et al. 2000). This information inspired the branch of computational cognitive studies known as connectionism (Rumelhart and McClelland 1986; McClelland et al. 1986). This approach suggests that learning occurs via the interaction of many components, which simultaneously constrain one another. Instead of a single processor, there are many processing elements. This architectural interpretation leads to the conclusion that knowledge is implicit in the structure of the device (brain or computer) as opposed to being explicit in a given state (Rumelhart 1998). Changing patterns of connectivity determine what is known and how the system responds to future input.

Connectionist models may provide answers to many classes of cognition problems. The chance for success is high because connectionist models have properties that align well with properties of human cognition. For example, the memory behavior of human experts exploits rapid pattern matching and completion, and rapid recognition response (DeGroot 1966; Chase and Simon 1973). Also, both connectionist models and humans detect correlations between actions and their outcomes, and adapt future behavior in response to this information. Connectionist systems allow defining learning procedures that permit the system to adapt to unforeseen stimuli (input), and exhibit a gradual adaptive "forgetting" behavior when unused connections gradually lose their influence on decision-making (output). Finally, connectionist networks learn via repeated training, which facilitates the study of the process of learning. The ability to study mistakes made during learning, and corrective adaptations to complex intellectual problems, opens up new and exciting avenues investigating how artificial and biological systems become experts.

There have been several noteworthy successes using ANNs to model human learning. Areas in which connectionist models have already expanded our understanding of cognition include the development of infant vision and perception (Bednar and Miikkulainen 2003; Chaput and Cohen 2001), language acquisition and comprehension (Elman 1991a; Miikkulainen 1997; Rumelhart and McClelland 1987), musical analysis (Todd and Loy 1991; Rumelhart 1998), and memory (Alvarez and Squire 1994; McClelland et al. 1995). A particularly interesting early connectionist model of learning was presented by Viscuso et al. (1989). Their ANN simulated qualitative reasoning while doing multiplication. This study is intriguing for many reasons. First, it successfully modeled how experts estimate correct answers. Second, by analyzing the type and frequency of errors during learning, Viscuso et al. (1989) showed similarities between their model and humans. Third, they successfully modeled association errors and showed that related data caused confusion, as it does with humans. The types of association errors they saw varied depending upon the order in which problems were learned and how much they were practiced. In summarizing their model, Viscuso et al. correctly pointed out that the most important contribution was the ability to mimic the manner in which experts rely not so much on formal logic and rules but on their "sense" of what is correct. Surprisingly, no studies have extended these results into other areas of mathematical learning and cognition. Since this study, significant advances have been made in technology and understanding of cognition, and it should be possible to give these results another look and to follow them up with studies that examine conceptual development, and learning more complex mathematical topics such as calculus.

Another interesting early ANN system by Cottrell and Tsung (1989) learned to perform arbitrarily long addition problems. Their model learned the implicit underlying rule of addition. They also analyzed the internal states of the model at different stages of learning and plotted the principal components of the results. They found that the network state space distinguished between actions needed to do addition: CARRY, WRITE, NEXT. Although the study focused mainly on rule use in problem-solving, this system showed the flexibility of ANNs for cognitive modeling: the network learned an important concept on which it had not been explicitly trained.

The ANN presented in this dissertation continues the application of connectionist systems to mathematics learning. The model advances previous work by expanding the context to include the effect of instructional delivery method on expertise learning. Because the computational studies in this dissertation are followed by a human subject experiment that investigates strategy development, the next section will discuss recent research on this topic.

### 2.2.2 Meta-cognition as reflected in Strategy Development

Several important research results increased interest in studies of meta-cognition. The most important result is that experts are highly meta-cognitive (Dörner and Schölkopf 1991; Scardamalia and Bereiter 1991). In fact, this is one of their primary attributes. Experts' problem-solving behavior does not simply follow a script, nor does it develop effortlessly. The greatest expertise results from long-term practice that is consciously goal directed, self-monitoring, and self-adjusting within the setting of each particular task (Ericsson 1998; Garner 1990; Hayes 1989). Therefore, to understand how learners become experts, we need to understand how they become meta-cognitive.

The development of meta-cognition is difficult to study directly, because it is such an abstract behavior. However, it is possible to monitor a learner's meta-cognitive development by studying their strategy development (Ericsson and Lehman 1996). The strategies that a learner chooses reflect how much she is consciously and successfully directing her own learning. When a learner creates increasingly successful strategies she has taken control of her learning, which means that she is being meta-cognitive, and the chances of acquiring expertise are very good.

Many psychological studies have investigated strategy choices as they depend on age (Siegler et al. 1996), memory (Cimbalo and Brink 1982) and conscious decision making (Siegler and Shipley 1995). Several of these results are useful for instruction. For example, studies have shown that meta-cognition is critical for successful academic learning (for a review see Sternberg (1998); Paris and Winograd (1990)). Also, children have difficulty laying aside old strategies, even when they are ineffective (Kuhn et al. 1992), and there is a natural human tendency to stop studying when a comfortable level of performance is reached (Ericsson 1998). These results imply that learners need help in overcoming a natural tendency to stop learning before reaching expertise. Fortunately, formal instruction influences how learners prioritize their study strategies (Siegler 1989). This means that there is an opportunity to influence learners and encourage intrinsic motivation.

Although the above results have improved our understanding of human learning, they have limitations as well. One such limitation occurs when studies focus on simple tasks, and on infants and children. This limitation is fine when the results are applied to early childhood development. However, intellectual expertise may begin many years later, and it may not be possible to generalize the results to adults in formal academic settings. Another limitation ocurrs when studies compare how groups of adults choose strategies without instructional intervention; the results therefore describe a stable cognitive state and not a developmental process. This limitation is a problem for researchers who want to understand how individual cognition responds to instruction. Altogether, there has been far less empirical research on strategy development of adults who succeed or fail to become intellectual experts. As a result, it is still unclear how meta-cognition, as reflected in strategy development, and instructional method interact.

The limitations of previous studies have several important implications for studies of expertise learning in formal education. First, learning tasks have to use realistic classroom topics. Otherwise the results may not transfer to an instructional setting. Second, studies need to focus on adult learners. Studies of infants and children may not apply to adults. Third, they need to analyze how learners respond to instructional methods. Instruction is a form of external intervention and we need to find out how learners perform in response to it. Such a focus will address a big instructional challenge, which is how to instill cognitive skills that support continued learning and improvement (Ericsson 1998).

## 2.3 Conclusions

The study of expertise has changed significantly since expertise was thought to be a general innate ability based upon superior heuristic search ability. Psychological studies of adult experts confirmed that expertise was learned and domain based. Today, much is known about general cognitive abilities common to experts. Most important of these is the ability to be meta-cognitive, a behavior that can be measured by monitoring strategy development. This understanding is encouraging because there is a role for educators in shaping the development of intellectual experts. Computational models have kept pace with these changes in the understanding of expertise, first reflecting the emphasis on search and more recently on developmental processes.

In spite of knowing a lot about experts, however, there is still much that we do not know about how expertise is achieved. In particular, we do not understand enough about how people can be helped to become experts through instruction. There is a role for both human studies and computational studies in striving to achieve this research goal. Continuing and expanding the historical trend of complementary human and computational studies of expertise, the following chapters will present research investigations based on three complementary methodologies: qualitative, computational, and quantitative psychological analysis. Together, this multi-method interdisciplinary approach provides more insight into expertise learning than could any one method alone. The foundation of this work, described next, is a qualitative research study that motivates both the development of a computational model and a human subject experiment about how instructional method affects learning for expertise in calculus.

# Chapter 3

# Research Questions

This chapter discusses a set of exploratory interviews with calculus experts that motivated the rest of the research presented in the dissertation. The interviews were designed to gain insight into the problem that although instructors are often experts in their field, it is not always easy for them to help their students become experts (Nathan et al. 2001; Nathan and Petrosino 2003). In particular, the interviews gathered information about the relationship between how instructors solve problems and how they teach students to solve problems. A data collection study was needed. Statistical studies were rejected for this exploration for two reasons. First, statistical studies are good for performing experimental manipulation, and not so good when a problem situation is ill-defined due to lack of data, as was the case here. Second, statistical studies are not the best way to collect data when there are no predefined response categories. In this investigation it was not at all clear what the experts would think and say. An exploratory study that could follow up on virtually any response by the experts was needed. Given these research needs, the exploratory interview was conducted using Grounded Theory. Grounded Theory is a qualitative research methodology that is designed to investigate phenomena without pre-determined hypotheses. Grounded Theory has the flexibility to pursue results in any direction. This chapter first introduces the interview study by describing the goals of the expert interviews. Next, an overview of Grounded Theory is provided. Next, the interview study is described, and the themes that emerged from the Grounded Theory analysis are presented. Then the chapter discusses how the interview results led to the development of three experimental hypotheses about how instructional method could affect teaching for expertise. Finally, the chapter concludes by discussing why a connectionist computational model is a good way to test the three hypotheses.

## 3.1 Motivation

In the last decade there has been an increasingly wide-spread national concern with properly educating citizens for the new technological age. There is widespread concern among academics that college students do not have critical thinking and generic problem solving skills. Mathematics and science achievement have been at the center of these concerns (for a few examples see Abudiab (2001); Medley (2000); Moore and Wick (1994)). Large-scale international studies have compared science and mathematics education in the United States with other countries (Atkin 1998; Stigler and Hiebert 1999). Far from resolving disagreement about which instructional methods to use, studies such as these have increased the debates. Often, reformers' opinions are based upon philosophical and political views (for three contrasting examples, see Eisenhart et al. (1996); Cromer (1997); Marshall and Tucker (1992).

Studies of expertise learning in mathematics can provide new insight about what instructional methods are most effective. We already have a lot of empirical data about expert cognition (material reviewed in Chapter 2). Now we need to learn more about how instructors translate their own expertise into instructional decisions. With that information, we can form empirically testable hypotheses about expertise learning. The results of testing these hypotheses may help improve instruction and learning in the many academic fields that depend upon calculus. So, the primary motivation for the study described in this chapter was to obtain enough information about the relationship between expert understanding and instruction in calculus, to be able to form empirically testable hypotheses.

An exploratory qualitative interview study was conducted because it is a good way to obtain information about a human phenomenon or problem that is not well understood, such as the relationship between expert understanding and instruction. In an interview, the researcher listens to what people themselves tell about their world, and learns about their views on their situation. In particular, a qualitative research interview attempts to understand the world from the interviewee's point of view, prior to scientific explanations (Kvale 1996). These attributes made a qualitative interview the best way to gather information about how the calculus experts perceive their expertise and their role in helping their students learn.

There were three goals for the interview study. Each of the goals addressed important aspects of expert self-perception, instruction and problem-solving. The discussion topic was calculus integration, because integration is one of the earliest classes that college mathematics students take and many advanced calculus topics build upon these early concepts. Another reason for choosing calculus integration was because there are well established methods of performing integration that students must learn (e.g. Simple integration, integration using Usubstitution, Integration by Parts). Instructors have learned to recognize the solution strategy for an integration problem intuitively, and they are motivated to help their students learn to identify problems intuitively as well.

The first interview goal was to determine how instructors categorize integration problems in terms of optimal solution strategy. This goal probed the instructor's thinking, with two desired outcomes. The first desired outcome was to find out how instructors decide what strategy to use. It was important to get this information directly from the experts so that incorrect assumptions about expert thinking would not taint future hypothesis development. The second desired outcome was to see how instructor decision correlated with their instructional methods. This information would provide insight about current practice; information that was needed before changes could be considered.

The second interview goal was to find out how instructors teach novice students to identify optimal solution strategies. This goal also gathered information about what current instructional practice for these experts was, but it also probed how much instructors thought about their practice. The desired outcome was two-fold. First, to look for the correlation between expert understanding and instructional method (as described in the first interview goal). Second, to find out if experts were meta-cognitive about this relationship. In other words, how much did instructors consciously think about how their own expert understanding affected their pedagogical decisions. This information is important because if calculus experts are also instructional experts they will be meta-cognitive about their instruction (Borko and Livingston 1989). If an instructor is not meta-cognitive about his or her instruction, this is something they can choose to change.

The third interview goal was to discover what problems instructors think novice students encounter when they attempt to categorize integration problems by solution strategy. The primary desired outcome of the third goal was to find out how much instructors were aware of the problems their students were having. The most important point of this goal was not to collect data about student problems (that information would be better obtained through a study of the students themselves). The most important point was to gather more information on expert meta-cognition about their instruction. An instructor who was meta-cognitive about instruction would be very aware of where her or his students were succeeding and failing (Borko and Livingston 1989) . This information is important, because instructors want to see their students succeed; if they know a lot about how students try to solve the same types of problems they themselves solve, they can more effectively help them.

The qualitative exploratory interviews were conducted using Grounded Theory. A Grounded Theory methodology is consistent with the needs of the expert interview study. The next section describes how research with Grounded Theory is conducted, so that the implementation and results of the interview study can be understood in their proper context.

## 3.2 Interviews with Experts

### 3.2.1 Overview of the Grounded Theory Method

There are several important methodological points that are core to how a Grounded Theory study is conducted. First, as with many other qualitative approaches to research, the Grounded Theory researcher goes to where a phenomenon of interest takes place. Laboratory experiments are uncommon in qualitative studies in general, because bringing someone into a lab removes them from their natural environment. When a participant (in this case an interviewee) is in their natural environment they are more comfortable and more likely to give a full account of their thoughts and lived experiences. In addition, the Grounded Theory researcher wants to observe the interviewee interacting with their natural environment, to gain additional perspective on their verbal reports. Second, controlled experiments are antithetical to Grounded Theory. Grounded Theory aims to understand human experience in all its depth and complexity. Controlling for any factor would necessarily invalidate the results. One way in which this philosophy appears in practice is that people are often referred to as “participants” rather than “subjects”. Third, a Grounded Theory investigation is often just one part of a larger ongoing study that may include other qualitative and quantitative investigations. As in this dissertation, sometimes a Grounded Theory investigation is conducted early on in a larger study, to provide insight into a poorly understood issue. Alternately, a Grounded Theory investigation may come in the midst of other research, in order to provide context for multiple distinct and unconnected sets of data. Finally, a successful Grounded Theory investigation always produces a highly contextualized understanding of the original topic of interest, and opens up new avenues for further research.

Unlike many statistically based research methods, which once developed mathematically rarely change, Grounded Theory has evolved over the years, much in the same way that a Grounded Theory study itself evolves. It is important to know a bit about this development, because it explains why there are some differences in how different publications describe the application of Grounded Theory. The seminal work by Glaser and Strauss (1967) spent much of its effort defending a general approach to social science research which rejected the scientific method. Less space was devoted to methodological procedure. This philosophical focus was a reflection of the time in which it was published. Behaviorism had made controlled experiments the norm in many disciplines, but there was a growing feeling among social scientists that this approach to studying human behavior was inadequate for describing social interactions. Before the authors could list procedures, they had to convince their audience of the need for their radically different approach to social science research. In particular, Glaser and Strauss continually reiterated that theory and hypotheses should not be formed prior to the start of data collection, because doing so restricted the possible outcomes of the analysis. Also, they emphasized the importance of widely open-ended data collection and continual mixing of analysis with data collection. This mixing allowed the researcher to identify and pursue important themes as they emerged during the study. However, the Grounded Theory researcher was continually reminded to resist forming conclusions prematurely or "fitting the data". Many people still refer to this original text to learn about Grounded Theory.

As Grounded Theory gained a strong following among social science researchers, and as qualitative research methods in general gained more acceptance, many researchers wanted more specific instructions on how to apply the theory. Recent books on Grounded Theory provide this information. Currently, the most cited book on the application of Grounded Theory is Strauss and Corbin (1998). They developed a coding-based methodology which provides rigorous guidance to the researcher, while building upon the original descriptions by Glaser and Strauss (1967). Strauss and Corbin's primary contribution is to stress that insight and understanding, and eventually theory, are obtained through strict application of coding procedures.

The expert interviews presented in this dissertation were analyzed using the Strauss and Corbin procedures for applying Grounded Theory. The following four sections will describe how the exploratory interviews with calculus experts were conducted, present the results and emergent theory, and discuss the implications for expertise learning.

### 3.2.2 Participants and Procedure

Structured interviews were conducted with two tenured Mathematics professors and one teaching assistant (TA). All three were currently affiliated with the University of Texas at Austin. The professors had extensive experience teaching introductory calculus, and the TA was a doctoral student whose research interests included Mathematics Education. All three interviewees will henceforth be referred to as "instructors". Each interview lasted 20-30 minutes, and the instructors were allowed to view the questions ahead of time (see Appendix A). There were 6 predetermined questions; each question addressed one of the three interview goals. The total number of questions was intentionally kept low in order to leave time for discussing additional topics brought up by the instructors. Questions were phrased to encourage descriptive responses. Additional questions were permitted when they followed up on topics brought up by the interviewees. The interviews were tape-recorded, transcribed verbatim, and analyzed for emergent themes using Grounded Theory.

### 3.2.3 Results

The instructors unanimously agreed that there exists no "algorithm" for selecting which method of integration applies to a given problem. One pointed out "with math there's always an exception to the rule". When asked to explain their solution selection strategy, none of the instructors directly addressed the question. Instead, they demonstrated how to perform the integration.

Although instructors could not explain how they categorized problems, their cognitive behavior was revealed when they answered a different interview question. Each interviewee was presented with the same set of four integration problems. The first three problems had similar surface-level features, but required different solution strategies. One problem was best solved using Simple Integration, another using USubstitution, another using Integration by Parts. The instructors agreed with each other about which of these optimal strategies to use. The fourth problem could not be solved using any of these strategies, and was included to stimulate additional conversation. Each instructor was then asked to take on the role of teacher explaining to a novice calculus student how they knew which solution strategy to apply. The request was phrased to emphasize that the task was not to solve the problem computationally, but to hit upon an optimal solution strategy on the first attempt. Instructors' initial reply often included phrases such as "it's obvious", "it's trivial" or "you just know this". One instructor referred to her ability to accurately categorize a problem as the equivalent of having a "large database of problems" in her head to which she could perform a comparison. However, she also said that she did not consciously compare problems to one another.

Everyone interviewed also agreed that an illusive ability to “just see it” is acquired through extensive practice. When pressed to speculate further, responses varied considerably. There were references to innate ability, being lucky, “thinking up something smart” and “just copying what other people have done over and over”. All the instructors believed that successful students spend extensive time with problems; one instructor even suggested that the learner needed to meditate upon problems. He was then moved to invoke the (by his own admission extreme) example of a famous mathematician who reputedly locked himself in an attic for seven years until he was able to arrive at the solution to a particularly difficult problem. Interestingly, one instructor commented that material is delivered so rapidly over the course of a semester that students do not have time to reflect on it. Two of the instructors also pointed out that their exams assess only computational ability, and not deep understanding. There was a contradiction between the instructors’ belief that novices need to spend lots of time absorbing rather than computing problems, and the way that the instructors presented material and assessed understanding. Oddly enough, although this pedagogical contradiction was acknowledged by one instructor during the course of conversation, neither he nor the others dwelled upon it. Given that thinking about this discrepancy might lead the instructors to make changes to help students learn better, it is puzzling that they did not appear to think it important.

All three instructors believed that learning should include deep rather than surface-level understanding. However, they had very different ways of trying to accomplishing this learning objective in the classroom. One instructor said that her lectures emphasized theory over computational skills. She explained that she taught this way because she thought it was the best way to acquire conceptual understanding of integration. Another instructor said that he emphasized explaining his own reasoning process to students. All of the instructors followed a traditional Drill and Test lecture format; they introduced integration topics sequentially, in isolation, with midterms and a comprehensive final exam. All of the instructors were often frustrated by how persistently shallow their students’ understanding was.

As one component of the interview, the instructors were asked to describe common student misconceptions. Surprisingly, the instructors were not confident about what kinds of cognitive problems their students experienced. Typical descriptions began with: “I would think”, “I hope”, "I don't know", "I would assume", and were punctuated by noticeable pauses. One instructor was convinced (and complained) that when faced with trigonometry, students automatically attempted the most difficult method of integration, and became tangled computationally in it. In contrast, another instructor categorically denied that students behaved in this way. She believed that students frequently avoided the more difficult integration solution strategies. Neither instructor had any concrete data to explain or support their beliefs; however, comments made during other parts of their interviews provide an intriguing clue. The first instructor claimed to teach students "when in doubt try parts" and the second faculty taught "always try substitution first".

### 3.2.4 An Emerging Theory of Instruction For Expertise

All three interview goals were met. The first goal, determining how instructors categorize integration problems by solution strategy, revealed that instructors categorize problems intuitively, but cannot explain how they do it. This result is important because it supports using categorization of integration problems as a topic for studying expertise learning. Specifically, the categorization problem fits the profile of an expertise problem in several ways. First, the instructors are calculus experts with a lot of domain knowledge. Second, they can rapidly and correctly categorize new calculus integration problems according to how they would solve them. Third, their categorization ability is intuitive and difficult to explain.

The first goal also successfully gathered information about how the experts' own abilities related to their teaching methods. The interviews showed that there was little relationship between how instructors thought about problems and how they instructed novices. This result is not surprising; because they could not clarify their own cognitive processes, it would be difficult for them to pass these processes on to novice learners. As a result, it is reasonable that the instructors would fall back on well-established methods of instruction such as Drill and Test.

The second goal, finding out how instructors teach novice students to identify optimal solution strategies, revealed that they use delivery methods that inhibit reflection and deep study, even though they know that these methods work poorly. As discussed in relation to the first goal, it is not surprising that the instructors relied on traditional lectures, and sequenced their concepts to follow textbooks. Without an alternative instructional approach, they stuck to what they knew.

The most interesting part of the second goal results, was the discovery that instructors were not meta-cognitive about their instruction. It seemed to bother the instructors that their students had so many problems learning to identify correct solution strategies; they were clearly aware of how this failing blocked many students from successful completion of the calculus course. On the other hand, they did not seem to take the next logical step, which would be to reflect on the situation, assess the factors which could contribute to the problem, and then experiment with instructional change. This lack of reflection implies that there is a need to develop alternative instructional delivery methods, test them, and present the results to interested mathematics faculty. This dissertation addresses this need through the experiments reported in future chapters.

The third goal, discovering what problems instructors thought novices encountered while categorizing problems, revealed that instructors have different opinions and little empirical data about student misconceptions. Their hesitation implied that they had not thought about student cognitive problems before. This helps explain why, in spite of having many years of teaching experience, they were not achieving expert results (defined as high levels of student success). The instructors were calculus experts but they were not instructional experts. Additional evidence supporting this conclusion is that the instructors were not displaying expert learning behavior; they were not immersing themselves in pedagogical issues, taking pedagogical risks, and learning from their failures. Instead, they were stalled and unsure how to improve their students' performance.

These results are consistent with the literature on expertise (material reviewed in Chapter 2), and lead to the following conclusions about instruction for expertise learning in calculus. First, instructional delivery methods that encourage students to develop intuition need to be used. Second, traditional delivery methods that isolate problems do not work, while immersion methods, that permit reflection, are impractical. Third, the delivery method should guide novices to spend more time with problems, and to search for deep structural relationships. Taken together, these conclusions support the theory that a delivery method that has structure, and also promotes comparing and contrasting problems for their structural relationships, will produce the best learning for expertise.

## 3.3 Development of Approach

### 3.3.1 Three Hypotheses to be Tested

As reported in the previous section, the theory developed from the expert interview data claims that a delivery method that has structure, and also promotes comparing and contrasting problems for their structural relationships, will produce the best learning for expertise. In order to simplify the following discussion, naming conventions need to be established for the delivery methods that will be tested experimentally. Drill and Test is a well known label in the educational literature, so it will continue to be used in this dissertation. Immersion learning will be referred to as Fully Integrated learning, in order to expand its use beyond the common association of "immersion" with foreign language learning. The new delivery method, proposed by the theory, will be referred to as Incremental Learning. This name reflects the use of structure and the gradually increasing complexity of this approach to instruction.

Three hypotheses will be developed, to test and compare the results of each of the delivery methods. Two of these hypotheses, concerning the effects of Drill and Test and Fully Integrated learning, have been developed directly from the interview study results. The third hypothesis is developed to tests the theory that emerged from analysis of the interviews.

*(1) Traditional sequential delivery methods inhibit learning and retention.* This hypothesis is clearly supported by the instructors interviewed, who stated that their Drill and Test teaching methods, which are commonly used, emphasized computational skills at the expense of deep understanding. Testing this hypothesis should confirm known results about human learning and provide a baseline to compare the other instructional delivery methods to.

*(2) Integrated delivery methods increase learning and retention.* This hypothesis is also clearly supported by the instructors, who say that in order to develop expertise, novices need to immerse themselves in material. Although Fully Integrated instruction is impractical to implement in the classroom, this hypothesis is important to test because it should confirm that the immersion approach has theoretical merit.

*(3) Incrementally increasing the complexity of the material will lead to the best performance.* This hypothesis directly tests the theory that emerged from the Grounded Theory analysis of the expert interviews. It should confirm that an Incremental Learning teaching method, one that is practical to implement in the classroom, will produce better expertise learning than either Drill and Test or Fully Integrated learning.

The previous section developed three hypotheses about expertise learning - if confirmed, they will provide new insight into improving instructional strategies.

### 3.3.2 Why a Connectionist Model?

Once the hypotheses have been established, the next step is to choose a good way to test them. Chapter 2 of this dissertation discussed why computational models are a good way to test hypotheses: simulations provide little risk to humans, and they can compress time. These advantages match the needs for testing the three hypotheses well. The performance of many students exposed to the three delivery methods needs to be compared. However, since Drill and Test results in poor learning, and Fully Integrated learning is impractical in the classroom, it is not desirable to expose classrooms of students to all three approaches for a long period of time, in order to compare their performance results. A computational model can address this issue. Simulations can include tests of these methods rapidly and without risk to live students.

The next methodological question is, what kind of computational model to use? Chapter 2 of this dissertation discussed the benefits of using a connectionist model for simulating human learning. In particular, connectionist models were shown appropriate for simulating developmental processes that are not rule-based. The data from the expert interviews revealed that the ability to identify solution strategies is intuitive. In particular, although the experts cannot explain their understanding, their comments strongly suggest that when they are exposed to new concepts (in this case calculus integration problems), their minds extract correlations between deep structural features. Thus their learning appears to be a statistical process. Connectionist models are statistical learning systems - they draw correlations based upon past experiences. Therefore, a connectionist model is an appropriate approach for studying expertise learning. The next chapter describes an artificial neural network that was used to test and confirm the three hypotheses.

## 3.4 Conclusions

The primary motivation for the interview study was to obtain enough information about the relationship between expert understanding and instruction in calculus integration, to be able to form empirically testable hypotheses. This goal was accomplished. Calculus experts were interviewed and the results showed that improved pedagogical delivery methods are needed to help calculus novices become experts. Current practice is not working well, and instructors are not sure what other instructional delivery methods would work better. A theory emerged from the interview data that the best student learning will come from a delivery method that has the benefits of traditional

Drill and Test learning (structure, ease of classroom application) and of immersion learning (reinforce comparison of problems and analysis of their deep structure), yet avoids the limitations of each method. Three hypotheses to test this theory were developed, and a connectionist model shown to be a good tool to test the hypotheses with. The next chapter will describe the connectionist model and experiments that tested the expertise learning theory.

# Chapter 4

# Testing the Hypotheses Computationally

The model used for the computational experiments in this dissertation is an artificial neural network (ANN). This chapter will describe the architecture of the model, and the data encoding of calculus integration problems. The ANN will be used to test the three hypotheses developed from the expert interviews. The design elements that apply to all computational experiments will be explained. Details of individual experiments and their results are reviewed, and the implications for expertise learning are discussed.

## 4.1 The Artificial Neural Network Model

### 4.1.1 Architecture and Data

The model is an artificial neural network trained with the backpropagation algorithm (Bishop 1995; Rumelhart et al. 1986) created using the LENS network simulator (Rohde 1999). The network is fully connected, and has 55 input nodes and 20 hidden nodes (Figure 4.1). The 55 input nodes make up a vector large enough to represent the features of one calculus integration problem containing up to four terms.

The input data consists of 957 calculus integration problems collected from three college level calculus textbooks, those of Lang (1986); Silverman (1985) and Stewart (1995). Feature coding is a logical choice for representing them, given that both novices and experts use the features

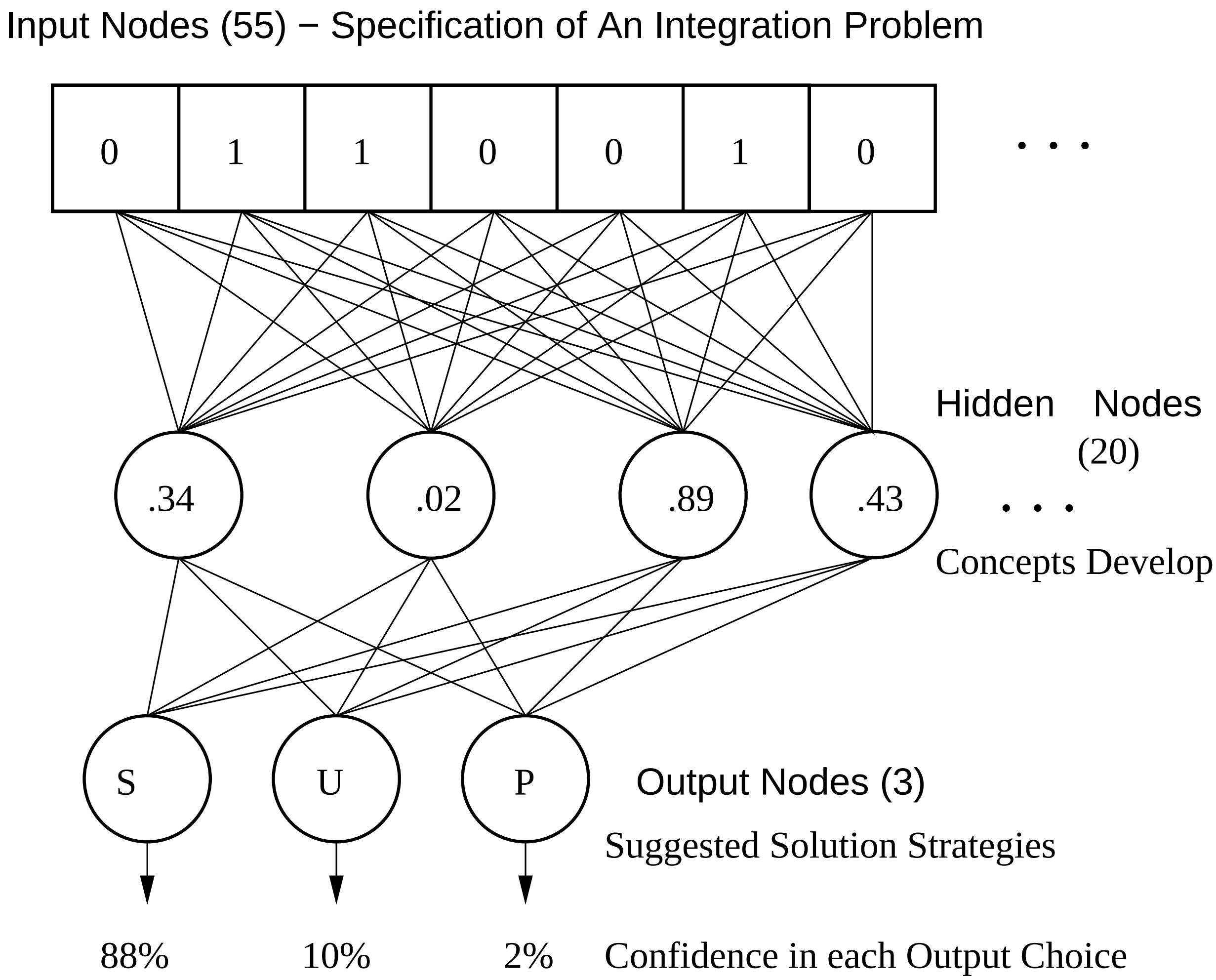



Figure 4.1: **The Computational Model.** A fully connected backpropagation neural network. Seven of the 55 input nodes are shown, 4 of the 20 hidden nodes, and all 3 output nodes.

of a problem to determine which solution approach to use (Chi et al. 1981). The 55 unit input vector contains a series of zeroes and ones that map operators and operands to their location in the calculus integration problem. A detailed description of the input vector is given in the next section.

The network has three output nodes, each of which represents one of the possible integration strategies: Simple, Usub, Parts. Because the network is trained with one active target at a time, it learns to represent how confident it is in each choice (Bourlard and Wellekens 1990). For example, if the network reports activation values at 88%, 10%, 2%, then it is quite confident in the first category, considers the second category possible but unlikely, and the third category extremely unlikely (but not absolutely impossible).

### 4.1.2 Description of the Input Vector

The 55-unit input vector is divided into the following parts:

- Four 2-unit groups representing constants and variables.
- Four 8-unit Unary Operators, representing $\sin, \cos, \tan, \cot, \sec, \csc, \ln$, exponentiation $e(x)$.
- Three 5-unit Binary Operators, representing multiplication, division, exponentiation ^, addition, subtraction.

In order to convert a human-readable integration problem into a form that can be input to the network, there is a 3-step coding process. The coding process will now be described. A running example, placed at the end of each step, will clarify the process.

***Step 1: Select an integration problem consisting of up to 4 terms.***

An integration problem may be shorter than four terms; the content must contain only the unary and binary operators listed above.

**Example**: $3 + cos(x) - sin(y) + ln(x)$

Term 1: 3
Term 2: cos(x)
Term 3: sin(y)
Term 4: ln(x)

***Step 2: Convert the integration problem to postfix format.***

Postfix format preserves the surface-level features of the problem, and puts them in an order that will be easier for input to the ANN.

**Table 4.1: Representation of Constants and Variables**

| Code | Variable Present? | Constant Present? |
|---|---|---|
| 00 | No | No |
| 10 | Yes | No |
| 01 | No | Yes |
| 11 | Invalid Code | |

**Example**: $3xcosysinxln + -+$

***Step 3: Convert the integration problem to a binary representation.***

***Step 3a: Code each term.***

Each of the four terms is coded into 10 units: 2 units for a constant or variable, and 8 units for a unary operator.

Format for a term: **00 00000000** (this example is blank)

**Constants and Variables:**

**00** 00000000

The first slot of every 2-unit group is reserved to indicate the presence of a variable, and the second slot is reserved to indicate the presence of a constant (Table 4.1).

Constants are not distinguished from one another. This reflects the assumption that calculus learners know that the important information is whether there is a constant or not in a particular location, not whether it is a 3 or 6 or some other number. This knowledge would have been acquired in previous mathematics classes. So constants are noted as present (1) or not present (0).

Variables are not distinguished from one another. This reflects the assumption that calculus learners know that the important information is whether there is a variable or not in a particular location, not whether it is an $x$ or $y$ or some other symbol. This knowledge would have been

**Table 4.2: Representation of Unary Operators**

| Code | Unary Operator |
|---|---|
| 10000000 | $sin$ |
| 01000000 | $cos$ |
| 00100000 | $tan$ |
| 00010000 | $cot$ |
| 00001000 | $sec$ |
| 00000100 | $csc$ |
| 00000010 | $ln$ |
| 00000001 | exponentiation $e(x)$ |
| 00000000 | No Unary operator |

acquired in previous mathematics classes. So variables are noted as present (1) or not present (0).

11 is an invalid encoding. The coding scheme does not cover the possibility of groups such as $3x$. Such a term contains an implicit operator (in this case multiplication) which would require additional slots to represent.

**Unary Operators:**

00 **00000000**

Each of the slots in every 8-unit group is reserved to indicate the presence of a unary operator (Table 4.2).

Unary operators are applied to either a constant or a variable. The encoding represents the presence of one of the eight operators: sin, cos, tan, cot, sec, csc, ln, exponentiation $e(x)$.

To review how each term is constructed, a full 10-unit term looks like this:

10 01000000 which in this case represents $cos(x)$

***Step 3b: Code each Binary Operator***

**Table 4.3: Representation of Binary Operators**

| Code | Binary Operator |
|---|---|
| 10000 | Multiplication |
| 01000 | Division |
| 00100 | Exponentiation ˆ. |
| 00010 | Addition. |
| 00001 | Subtraction. |
| 00000 | No binary operator. |

Each of the three binary operators is coded into 5 units.

Format for a binary operator: **00000** (this example is blank)

Each of the slots for the three 5-unit operators is reserved to indicate the presence of one of the Binary operators: multiplication, division, exponentiation ˆ, addition, subtraction (Table 4.3)

**Example**: Using the above description of encoding, our postfix integration problem is encoded like this:

01 00000000 10 01000000 00 10000000 10 00000010 00010 00001 00010

where the components are:

Term 1: 01 00000000

01 : Constant (i.e. 3)

00000000 : NONE (i.e. no unary operator for the constant)

Term 2: 10 01000000

10 : Variable (i.e. $x$)

01000000 : cos (of the variable $x$)

Term 3: 10 1000000

10 : Variable (i.e. $y$)

10000000 : sin (of the variable $y$)

Term 4: 10 00000010

10 : Variable (i.e. $x$)

00000010 : ln (of the variable $x$)

Binary Operator 1: 00010 : +

Binary Operator 2: 00001 : -

Binary Operator 3: 00010 : +

A clustering analysis, conducted using Principal Component Analysis, shows the distribution of the coded integration problems (Figure 4.2).

### 4.1.3 Experimental Design

The calculus integration problems were divided into 10-fold cross-validation training and test sets (called splits, or learning experiments). In each experiment the training set was input to the network, one problem at a time, in random order, and the test set was used to measure performance. Validation sets were not used to decide when to stop training because each learning experiment represented training one subject and the training time had to be constant, to compare how well the subjects learned. Three different types of learning experiments were run. Each experiment was run ten times, randomly resetting the initial network weights each time. Thus the whole study consisted of 300 learning experiments. This way it was possible to model the behavior of many different subjects and measure both emergent patterns and individual variation.

During the test phase, there was always only one correct answer to a problem. This answer, called the "Best", was the answer suggested in a textbook, or by a calculus expert (mathematics faculty or TA).

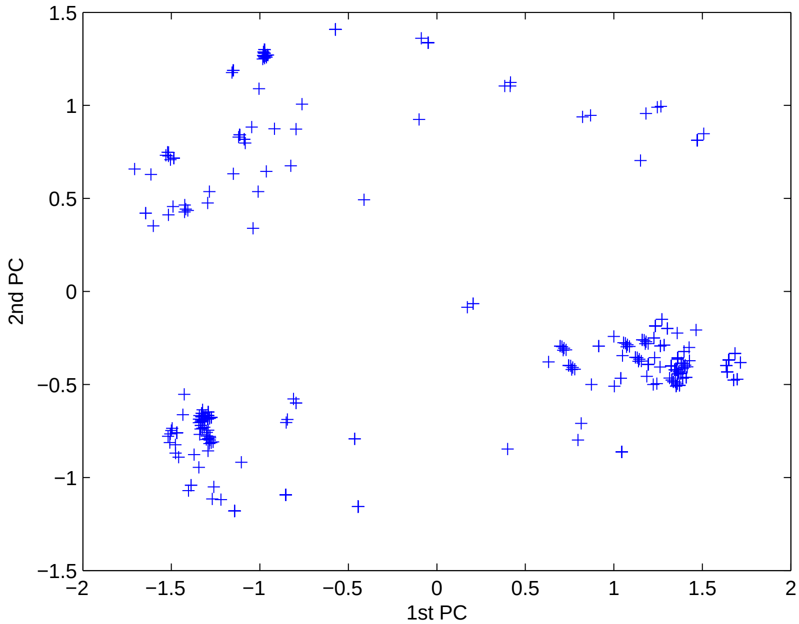



Figure 4.2: **The Integration Problems.** The distribution of the 957 coded calculus integration problems, created using Principal Component Analysis.

## 4.2 Computational Experiments

The three hypotheses will be tested computationally in this section. Each of the hypotheses tests learning and performance under training with a different delivery method: Drill and Test, Fully Integrated, or Incremental Learning. The expert interviews provided a learning task that could be used for all the experiments. The calculus experts could identify solution strategies for integration problems, but novices could not. Novices need to acquire this ability in order to solve integration problems under tight time constraints (like exams). Therefore, the problem chosen for the simulations is to decide whether a given integration problem should be solved with Simple Integration (Simple), Integration by USubstitution (Usub), or Integration by Parts (Parts). The following sections will simulate learning this problem using Drill and Test learning, Fully Integrated learning, and Incremental learning. This chapter concludes with an analysis of the three sets of simulations and how they support the hypotheses developed from the expert interviews.

### 4.2.1 Drill and Test Learning Simulations

The first set of experiments, called "Drill and Test", mimicked a classic form of delivery that results in poor long-term retention and conceptual understanding in humans (Carpenter et al. 1980; Resnick and Ford 1981; Ross 1988). In this method, concepts are introduced to the learner one at a time, with no overlap between topics. At the end of each topic, the learner is given a midterm exam (of previously unseen examples) on that concept. Prior to taking a final comprehensive exam the learner has a short "cram session", i.e. all concepts are trained on at the same time.

In order to monitor the progress of learning quantitatively, and to compare to other approaches, each network was also tested during each epoch in two ways: (1) with the current midterm exam, illustrating the performance that the teacher would see in the classroom (Figure 4.3), and (2) with the comprehensive exam, monitoring progress in learning the entire task; the results were broken into separate statistics for each concept (Figure 4.4).

The main result was that the model, like humans, only remembers the most recently introduced concept well. More specifically, in 100 experiments run using Drill and Test, most networks (83%) rapidly learned to identify each of the concepts in turn (Simple, Usub, Parts; Figure 4.4). However, in spite of the opportunity to cram first, when the comprehensive final exam was given, these learners performed poorly, averaging 41.65% (standard deviation 6.35; Figure 4.3). The median score was similar, at 42.31% (interquartile range 10.54). The highest score was 54.55%. These scores, somewhat above chance, occurred because the learners were still choosing Parts as their answer most of the time. The remaining 17% of network learners were unable to make the switch from Simple problems to Usub or Parts problems. Their scores remained flat (0%) during the Usub and Parts training periods. The cram session had a small beneficial effect for these learners. However, their final scores were roughly equivalent to random guessing, averaging 17.29% (standard deviation 4.95), with a high score of 26.92%. The median score was similar to the mean, at 16.00% (interquartile range 8.1).

### 4.2.2 Fully Integrated Learning Simulations

A second set of experiments mimicked human learning using an approach called "Fully Integrated Learning". This approach is inspired by the immersion experiences popular in foreign language learning (Spolsky 1989): The learner is placed in an environment where she or he is completely surrounded by the stimuli to be learned. In the Fully Integrated Learning experiments, there was only one training period, during which the networks were trained on all of the problem types simultaneously. During each epoch, the Simple, Usub and Parts training problems were input to the network in random order. Exams using the entire test set were given after every training epoch. Time is compressed in the Fully Integrated simulations, to focus on the learning process.

Fully Integrated Learning produced significantly better results than the Drill and Test delivery experiments ($t = -9.240, df = 16.938, p = 5.015e - 08$). The average score on the final comprehensive exam was 76.99% (standard deviation 7.94; Figure 4.5). The median score was similar to the mean, at 76.92% (interquartile range 11.33). The highest score was 80.76%. The errors that were made on the exams followed a pattern of slow, gradual learning, spread across all problem types (Figure 4.6). The Fully Integrated learning experiments as a whole replicated human data showing that immersion leads to better long-term retention than does Drill and Test.

### 4.2.3 Incremental Learning Simulations

The third set of experiments was designed to test the hypothesis that the best learning of material is obtained by Incremental Learning. This approach is inspired by the result in the machine learning community that it is often most effective to tackle large computational tasks by starting with small problems and gradually increasing their complexity (Elman 1991b; Gomez and Miikkulainen

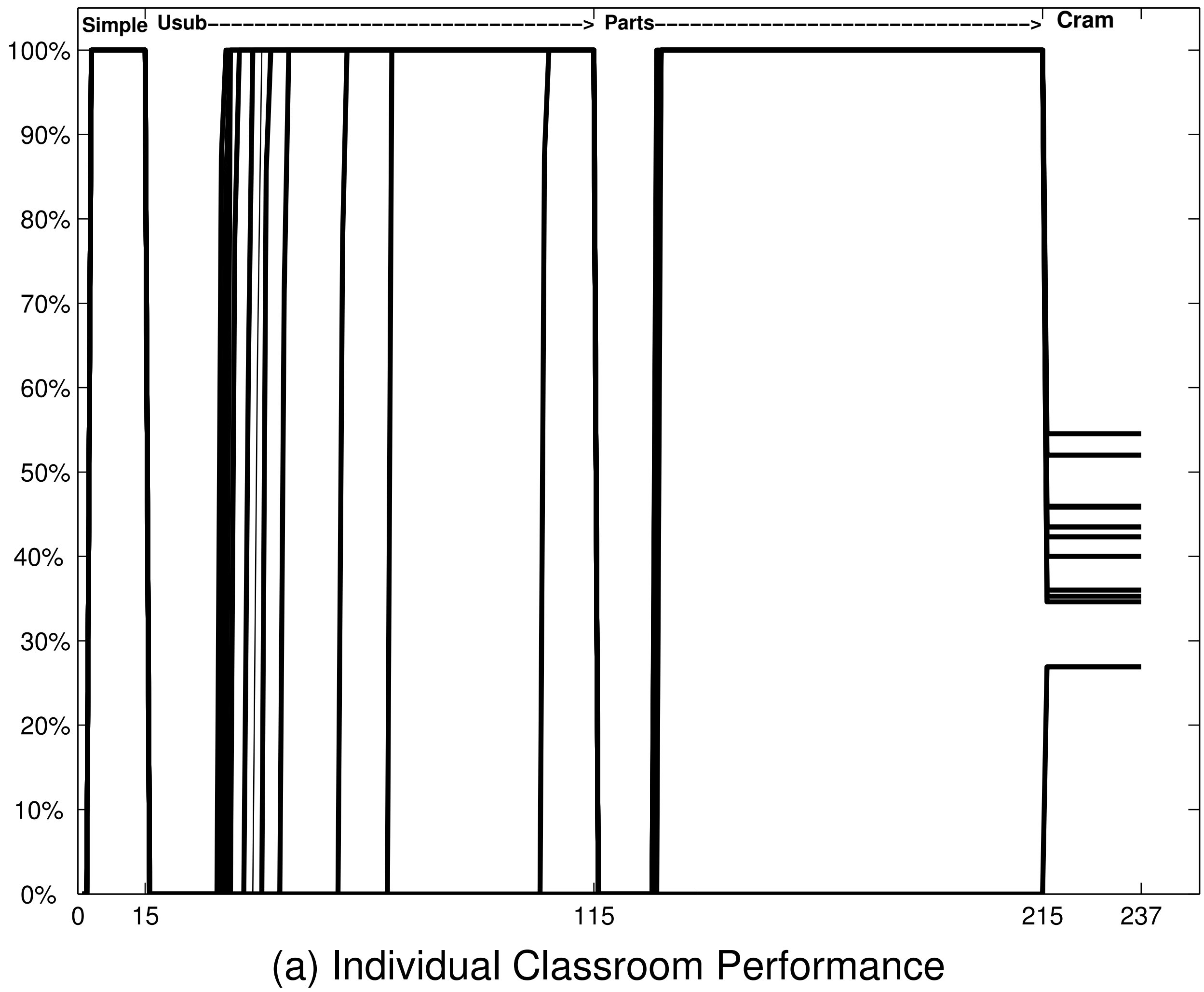



Figure 4.3: **Drill and Test Learning Performance.** The performance of 12 representative network learners is shown as they are trained to categorize integration problems, one concept at at time. Exam scores are on the y-axis, and the training epoch is shown along the x-axis. Simple integration problems are introduced for the first 15 epochs; the learners learn to recognize them almost immediately, and exam scores shoot up to 100%. When Usub problems are introduced, and Simple problems removed, at epoch 15, the learners fail to recognize Usub; exam scores drop sharply to 0%. One at a time, most of the learners make the sudden shift to recognizing Usub as the correct categorization; exam scores rapidly rise to 100%. When Parts problems are introduced, and Usub problems removed, at epoch 115, the same pattern repeats. i.e. all the learners immediately fail to categorize the problems correctly and their exam scores drop to 0%. Shortly thereafter, most of the learners make the switch to selecting Parts and their scores shoot up to 100%. The cram session begins at epoch 215; Simple, Usub and Parts problems are trained all together, and the exams are comprehensive. This mixed training has a small beneficial effect for the poorest learners; their average score rises to 17.29% (maximum score 26.92%). Conversely, the cram session reduces the scores of the learners who had scored well on Parts exams. Their final average score drops to 41.65% (maximum score 54.55%). In sum, all of the learners trained using Drill and Test performed poorly on the final comprehensive exam (epoch 237), because they could not distinguish the different concepts from one another.

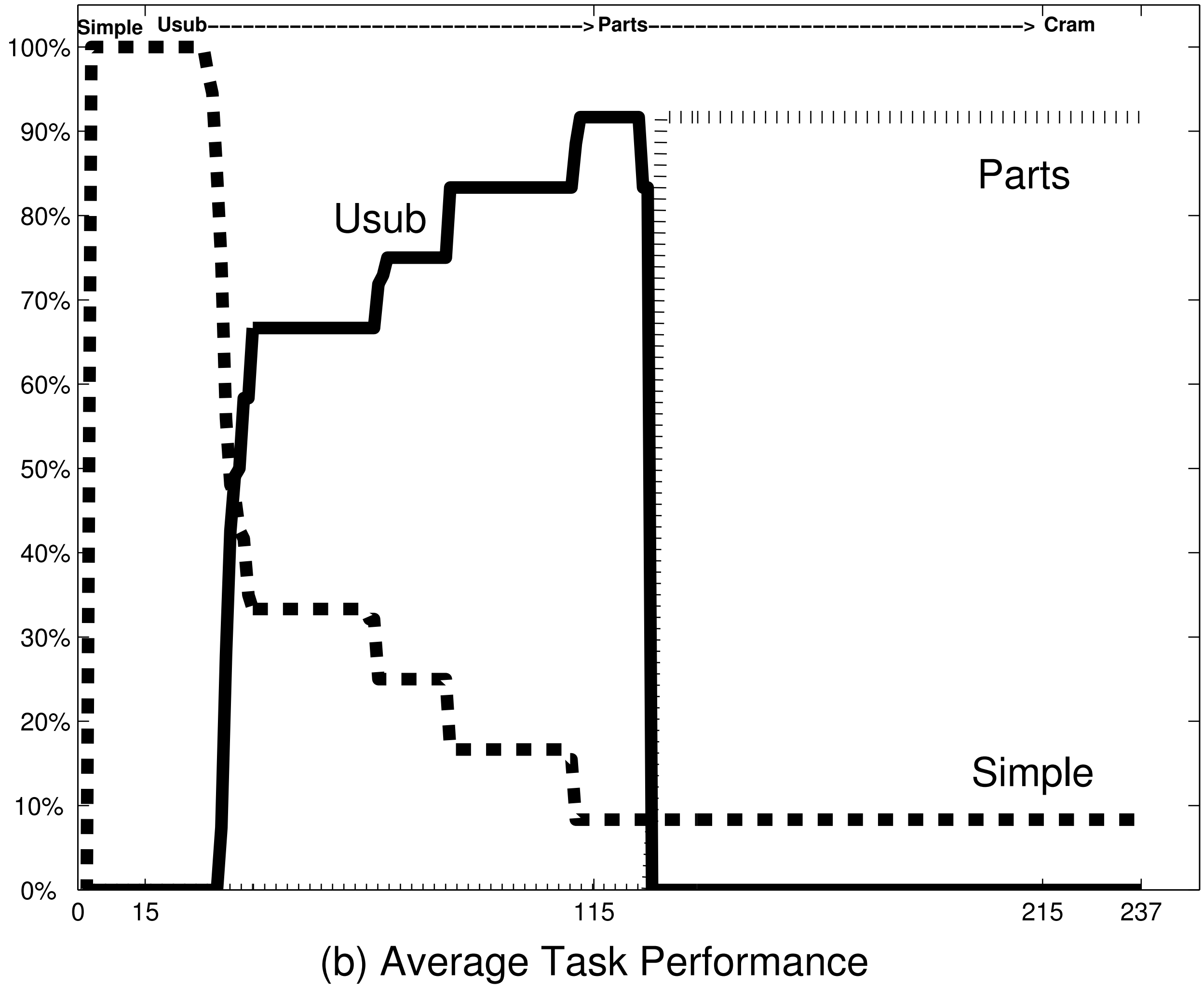



Figure 4.4: **Drill and Test Learning by Concept.** Average performance of all 100 learners broken down by concept (Simple, Usub, Parts). This figure shows how well each learner performs on a comprehensive exam taken during each epoch of training. As would be expected after viewing the data in Figure 4.3, the learners are able to recognize Simple problems immediately - their average score shoots immediately to 100% (the thick dashed line). Average scores for Usub and Parts are 0%, because the learners have not been introduced to these concepts. When Usub problems replace the Simple problems (epoch 15), most of the learners eventually learn to recognize them (the solid black line). The uneven ascent of the Usub line, matched with the uneven descent of the Simple line, occurs because when learners begin to recognize Usub problems they forget how to recognize Simple problems. The Usub line never reaches 100% because some learners never make the switch, pulling down the average score. The Parts scores remain at 0%, as expected. When Parts problems replace Usub problems (epoch 115) the average score for Parts shoots up (the narrow dashed line). The Parts line never reaches 100% because not all learners learn to recognize Parts problems. Some learners continue to choose Simple, which explains why the Simple line remains around 10%. The Usub problems drop to 0%. Given that Usub problems took longer to learn than either Parts or Simple problems, this sharp drop to 0% suggests that Usub problems are the hardest concept to learn and retain. This figure complements Figure 4.3, suggesting that each concept is forgotten when a new concept is introduced.

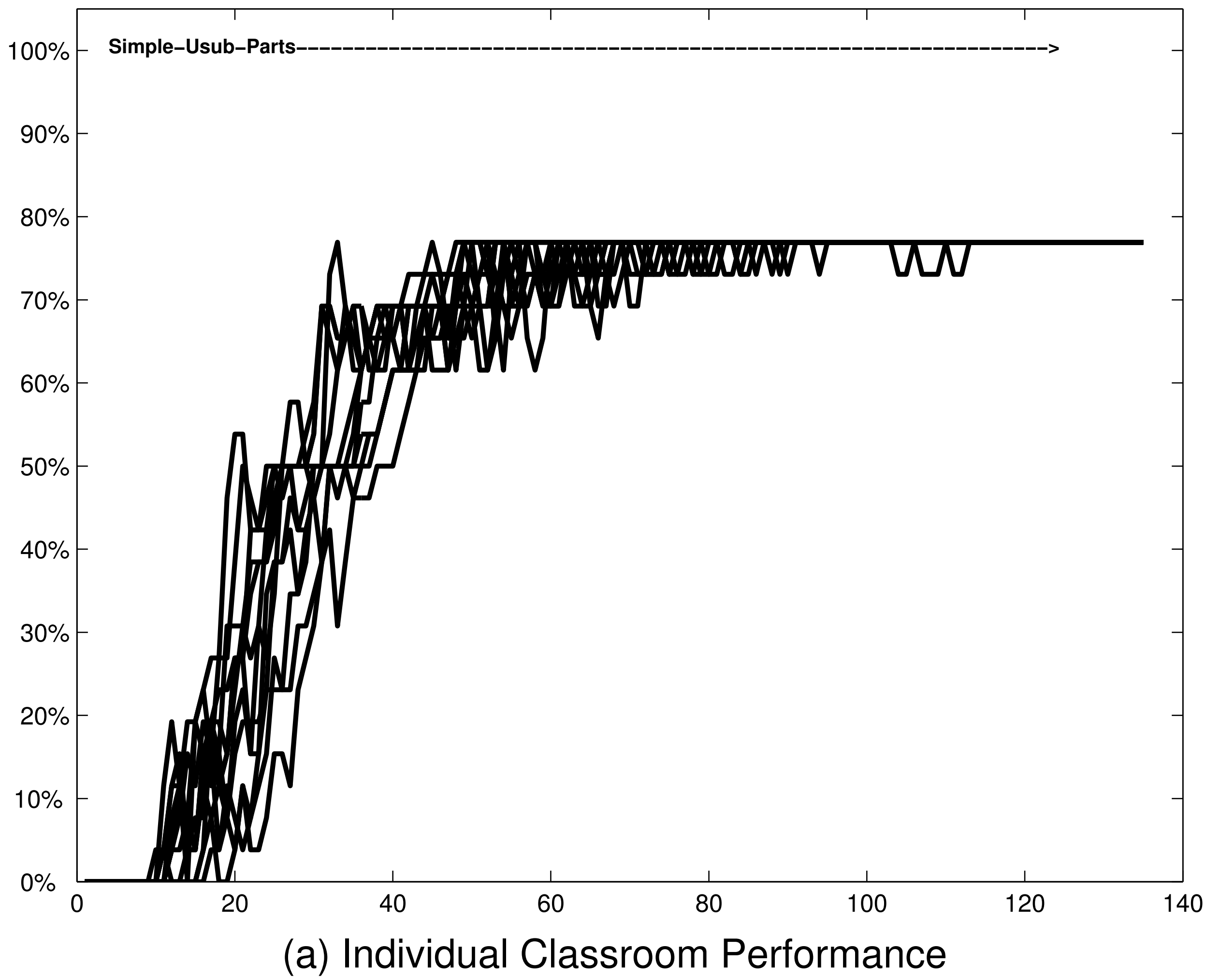



Figure 4.5: **Fully Integrated Learning Performance.** The performance of 12 representative learners as they are trained to categorize integration problems, all three concepts at the same time. Exam scores are on the y-axis, and the training epoch is shown along the x-axis. For approximately the first 10 epochs, all the learners score poorly. Most network learners score 0% (shown here), and others score slightly better as a result of successful random guessing (not shown here). Then, they begin to improve; by epoch 20 all of the learners are scoring above 0%. Each of the learners rapidly improves at correctly categorizing the concepts. Their progress is not smooth, reflecting the trial and error process of learning. Eventually, the learners' performance plateaus at approximately epoch 50, with an average exam score of 76.99% (high score 80.76%). None of the learners achieve perfect exam scores, no matter how long training continues. In sum, all of the Fully Incremental Learners perform well.

1997). An Incremental Learning delivery introduces new, increasingly complex concepts along with reinforcement of old concepts.

As with the Drill and Test experiments, there were three training periods. The network was first trained to identify Simple problems. During the second training period, Usub problems were added to the Simple problems, and for the third training period, Parts problems were added. The

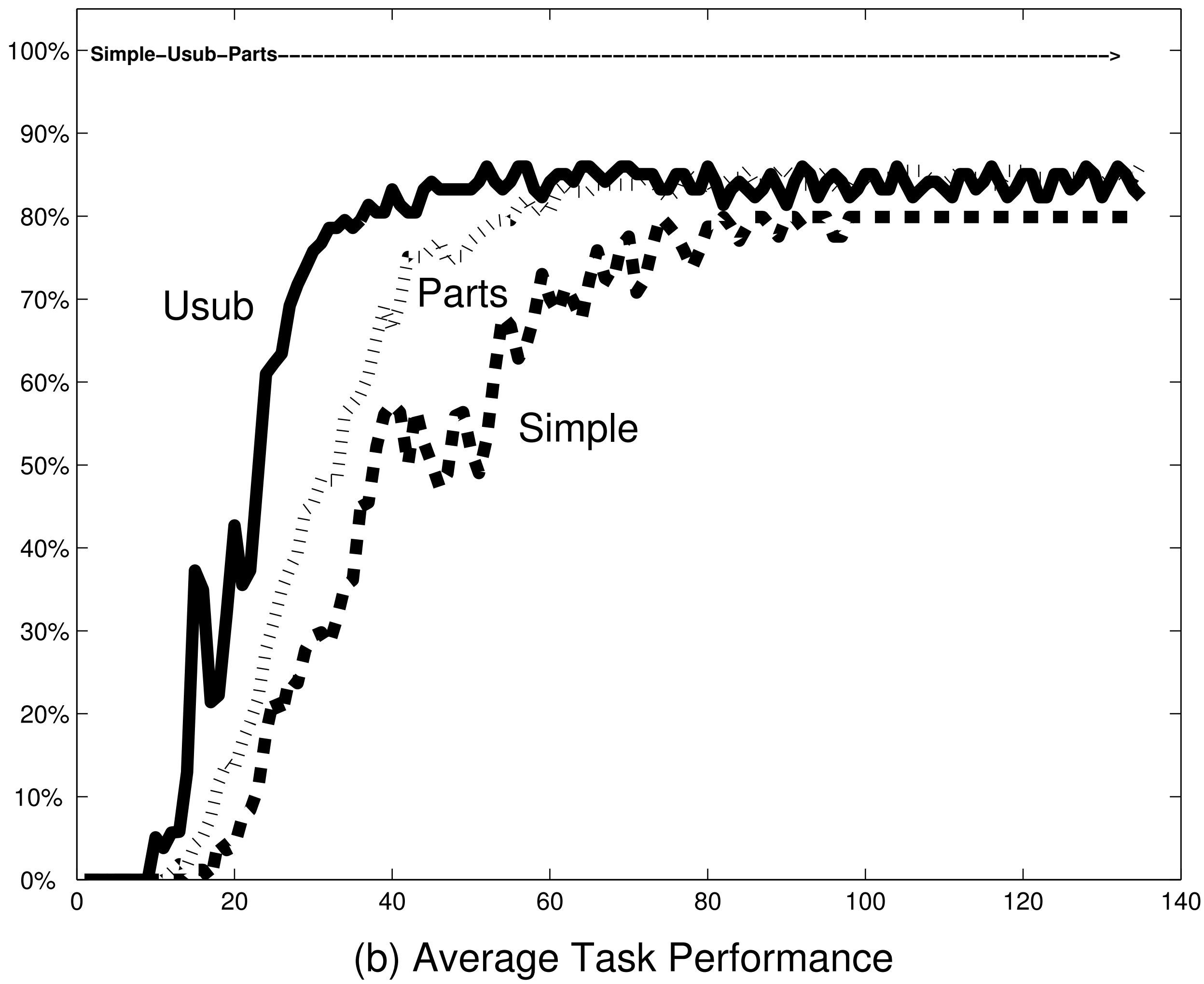



Figure 4.6: **Fully Integrated Learning by Concept.** Average performance of all 100 learners broken down by concept (Simple, Usub, Parts). This figure shows how well each learner performs on a comprehensive exam taken during each epoch of training. The results show that each of the three concepts is learned at a similar rate, at about the same time, and that at the end of training the learners are able to correctly categorize all of the concepts equally well. These results contrast sharply with the results for Drill and Test (Figures 4.3 and 4.4), by showing that Fully Integrated learners appear to have a good understanding of all three concepts.

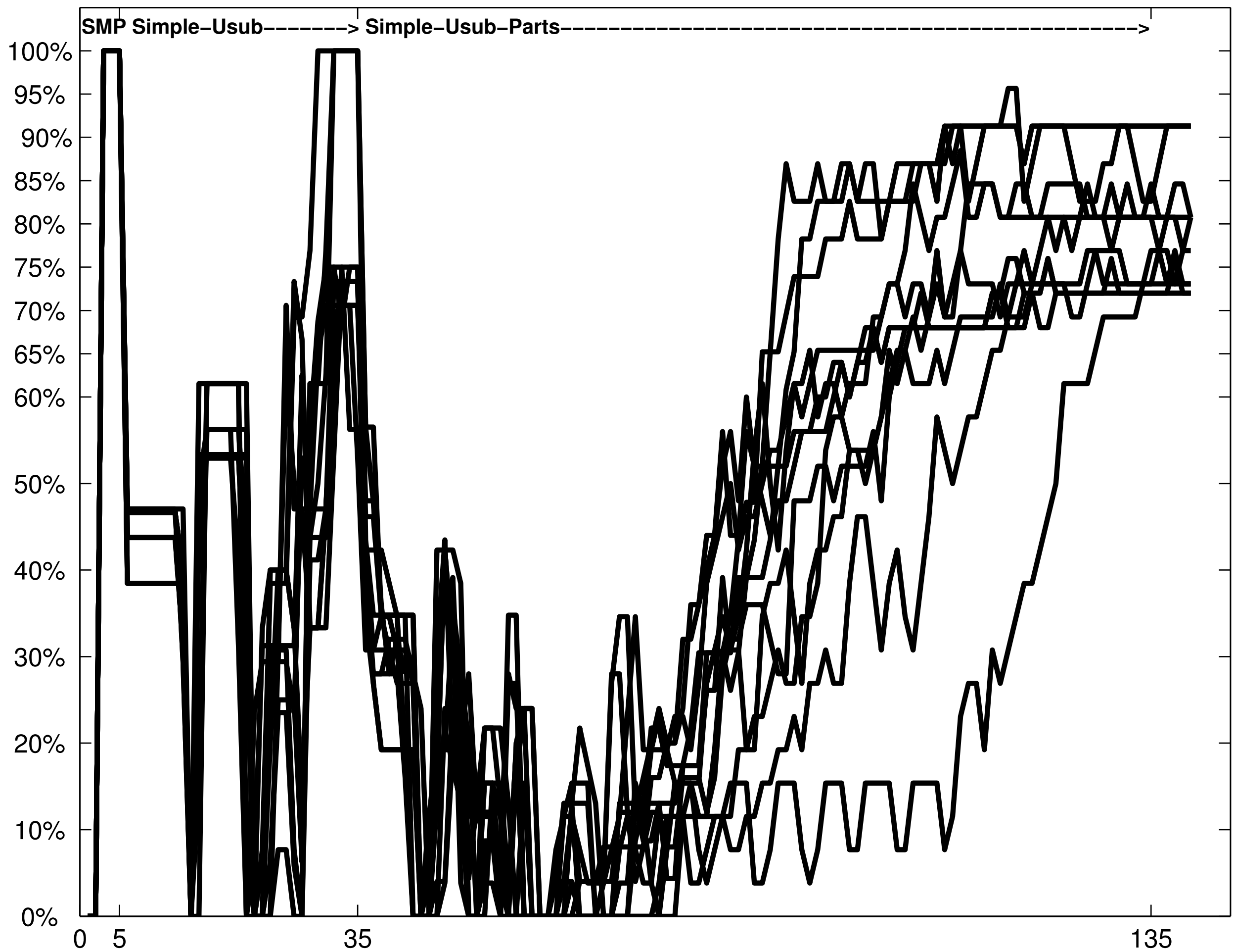
classroom performance was measured with Simple tests during the first period, Simple and Usub test problems during the second, and the entire test set during the third (Figure 4.7). The progress in learning the entire task was monitored with the entire test set, broken down by concept (Figure 4.8). As in the Drill and Test experiments, the network rapidly learned to identify Simple problems. When Usub problems were introduced, test scores began to fluctuate severely. Over time, although fluctuation continued, overall test scores increased. When Parts problems were introduced, the pattern of fluctuating scores was accentuated. Midterm scores immediately plummeted, although it is interesting to note that even the downward drop was often not smooth, but marked by brief plateaus and recoveries. Eventually, performance began to improve, with prominent individual differences, as each network learned subtle patterns to accurately identify each concept. The average score on the final comprehensive exam was 81.9% (standard deviation 8.23). The median score was 78.85% (interquartile range 10.6). It is important to note that the final test results for Incremental Learning were better than either Drill and Test or Fully Integrated Learning, in spite of interim results that sometimes appeared poorer than either other type of experiment. The maximum comprehensive exam score was 95.6%, higher than any score reached in a Fully Integrated learning experiment. As evaluated with a t-test, the Incremental Learning final exam scores were higher than those of the Fully Integrated learning ($t = 1.957, df = 11.869, p = 0.074$).

## 4.3 Discussion

The computational experiments validated the general approach of learning to categorize calculus integration problems by their solution strategy. Two delivery methods (Drill and Test, Fully Integrated) mimicked known data about human learning, and the results matched results reported in the expert interviews. The three hypotheses of expertise learning developed from the expert interviews were confirmed.

*(1) Traditional sequential delivery methods inhibit learning and retention.* The Drill and Test network learners retained material poorly in the long term. Their good scores during training on isolated concepts implied understanding, yet they performed poorly when comprehensive exams forced them to distinguish one concept from another. This result suggests that Drill and Test learners memorize concepts in the short term and later forget them. This suggestion matches the literature on traditional instruction (discussed in chapters 1 and 2) and the expert interview reports that human calculus students taught with Drill and Test perform poorly and have shallow understanding.

*(2) Integrated delivery methods increase learning and retention.* Fully Integrated network learners, who are immersed in all the integration concepts at once, took longer, but eventually learned better than Drill and Test learners. This result suggests that at first it is nearly impossible to distinguish multiple complex concepts; this confusion lessens given sufficient time to compare and contrast problem examples. This result matches the literature on immersion learning discussed




(a) Individual Classroom Performance

Figure 4.7: **Incremental Learning Performance.** The performance of 12 representative learners as they are trained to categorize integration problems. Exam scores are on the y-axis, and the training epoch is shown along the x-axis. Simple integration problems are introduced for the first 5 epochs; the network learners learn to recognize them almost immediately, and exam scores shoot up to 100%. When training on Usub problems is added to training on Simple problems at epoch 5, scores immediately drop. A pattern of sharp improvement followed by sharp failure continues for several epochs, however the peaks gradually increase. At epoch 35, a few of the learners achieve scores of 100%; most of the learners reach peaks in the mid 70s, and show a pattern of overall improvement. At epoch 35, training on Parts problems is added to Simple and Usub problems. The same pattern occurs as previously seen: there is a sharp drop in exam scores, broken by a plateau; this is followed by sharp fluctuations in scores. This fluctuation lasts longer than before, and for a while the peak scores become lower. Although performance scores are poor, the amount of fluctuation is decreasing. Eventually, each of the learners begin to improve, just as they had in the Simple-Usub training segment. Improvement is highly individualistic, with some learners improving faster than others. Eventually, all the learners plateau, at an average score of 81.9% (high score 95.6% at epoch 135). In sum, Incremental Learners perform better than either Drill and Test or Fully Integrated learners over the long term, and appear to have the best understanding of the concepts.

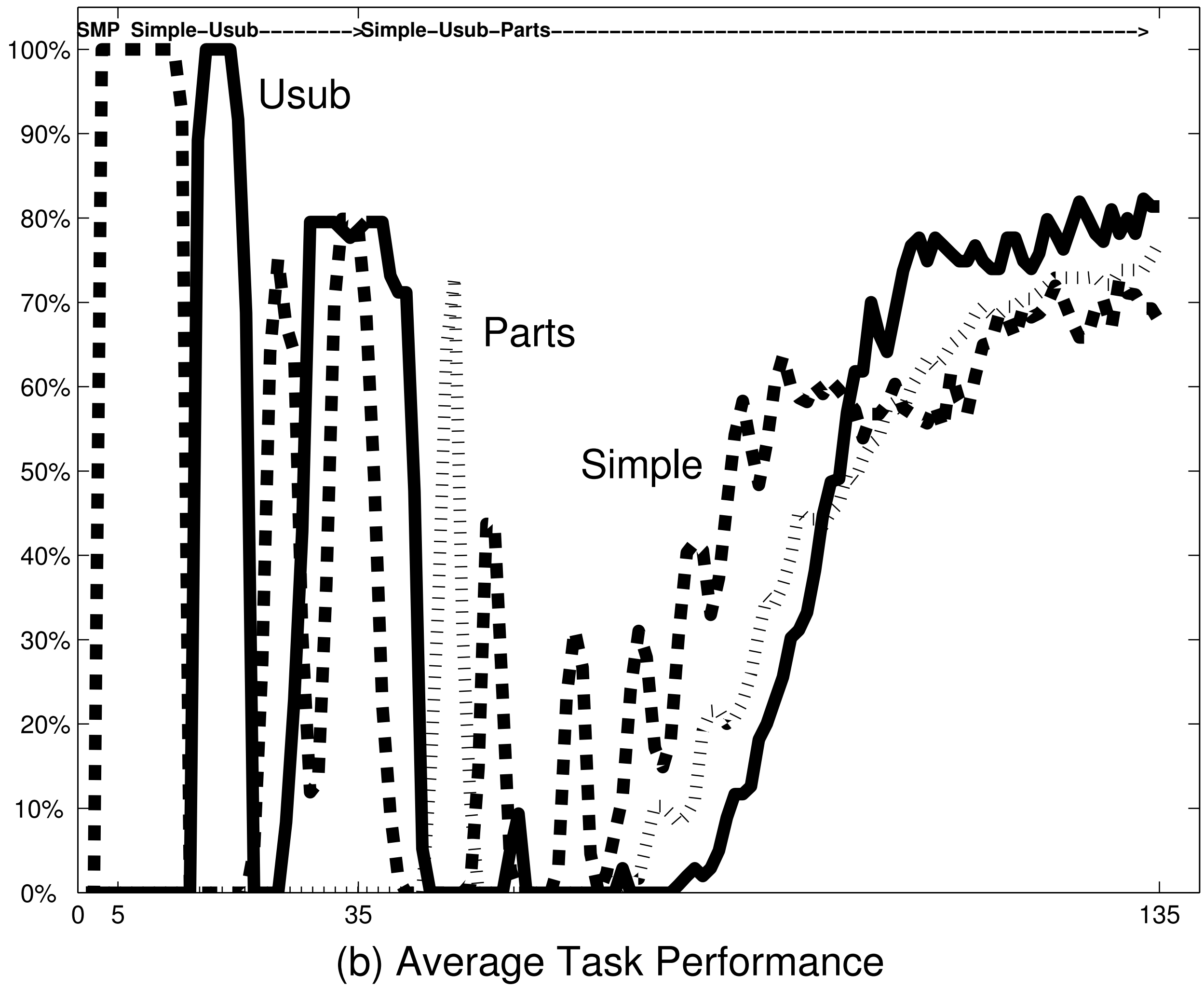



Figure 4.8: **Incremental Learning by Concept.** Average performance of all 100 learners broken down by concept (Simple, Usub, Parts). This figure shows how well each learner performs on a comprehensive exam taken during each epoch of training. As would be expected after viewing the data in Figure 4.7, the learners are able to recognize Simple problems immediately - their average score shoots immediately to 100% (the thick dashed line). Average scores for Usub and Parts are 0%, because the learners have not been introduced to these concepts. When Usub is added to training at epoch 5, Simple and Usub line (the solid black line) alternate performance: when Usub is at 100%, Simple is at 0%. This implies that the learners alternate between recognizing one concept or the other. Gradually, the Simple and Usub lines begin to approach each other: neither scores perfectly, neither fails completely. At epoch 35, the two lines come close to each other in the high 70s. This implies that the learners have learned to distinguish the two concepts from one another fairly well. When Parts problems are added to training at epoch 35, there is a similar set of behaviors. At first, Simple and Parts lines drop together to 0%, while Parts (the narrow dashed line) rises sharply. Then, the Parts line drops sharply, while the Simple line rises. Following that, the Usub line rises just a bit from zero, while the Simple line joins Parts at 0%. This implies that, as previously, the network at first alternates between recognizing each of the concepts. Finally, all three lines rise, and by the end of training at epoch 135, there is little fluctuation. This implies that the network is now able to distinguish all three concepts well.

earlier, and the expert interview reports that deep understanding requires a lot of time spent "meditating" on the material.

*(3) Incrementally increasing the complexity of the material will lead to the best performance.* Incremental Learning resulted in the best long term retention of material and the best learning performance. Learners struggled to learn the concepts and appeared to be comparing and contrasting them, prior to being able to distinguish one from another. As each new concept was added, the period of struggle got longer, and interim performance appeared poorer. Learners follow very individualized learning paths, but in the long run, Incremental Learning always resulted in the best average performance and the highest individual performance. These results imply that Incremental Learning is a good delivery method to try with human learners. Incremental Learning combines the advantages of Drill and Test (structured delivery) with the advantages of Fully Integrated (forced comparison of concepts), and avoids the limitations of both delivery methods.

## 4.4 Conclusions and a Prediction

The computational experiments supported three hypotheses generated by the expert interviews: (1) Traditional sequential delivery methods inhibit learning and retention, (2) Integrated delivery methods increase learning and retention, (3) Incrementally increasing the complexity of the material will lead to the best performance. These conclusions suggest that there might be a better way to instruct human learners to become experts. Of course, expertise develops over a period of several years, and formal learning environments, such as classrooms, can not compress years into weeks or months. However, experts have meta-cognitive abilities that are distinct and different from those of non-experts. One of these abilities, effective study and test taking strategies, can be acquired fairly quickly by a motivated novice. An Incremental Learning delivery should encourage the development of effective study and test-taking strategies. When this happens, the learner stands a greater chance of eventually becoming an expert. The next logical step, therefore, is to test this prediction on human subjects. The following section describes a study where the effect of each delivery method was compared on human performance, strategy development, and conceptual development.

# Chapter 5

# Testing the Computational Prediction with Human Subjects

This chapter discusses an empirical study designed to test the prediction developed from the computational experiments, that an Incremental Learning delivery method will encourage meta-cognitive abilities necessary to achieve expertise. In particular, the Incremental Learning subjects' study and test-taking strategies, conceptual development, and performance were evaluated to see if they were more effective than those developed by Drill and Test or Fully Integrated learners. The experiment consisted of a formal laboratory study followed by a structured interview. The study was designed and analyzed using Meaning Categorization, an interdisciplinary research methodology. Qualitative and quantitative results about strategy development, conceptual development and performance show that an Incremental delivery produces the best environment for expertise learning. Unexpected results about the affective reactions of participants show that only the Incremental Learners responded positively to the learning task. Finally, the chapter discusses how the human subject results support the prediction that Incremental Learning is a better instructional delivery method for expertise learning.

## 5.1 Motivation

The expert interviews established that calculus instructors are often frustrated when their students have trouble learning integration concepts. The instructors report that many of their students have only a surface level understanding of problems and so can not choose correct solution strategies. The interviews also established that instructors, who are calculus experts, have trouble explaining how they correctly identify solution strategies themselves. Instructors rely on an intuitive understanding that developed gradually after many years of study. Perhaps not surprisingly, given that they can not explain their own cognitive processes, the instructors teach their classes using well established delivery methods such as Drill and Test. The instructors know that this approach does not work well, and would like to find out about a more effective delivery method. They believe that a long-term immersion would be better for expertise learning than Drill and Test, but acknowledge that this approach is not practical for the college classroom. Instead, they would prefer an instructional approach that has the advantages of structure and immersion.

In response to these needs, a theory was developed in the previous two chapters that a delivery method called Incremental Learning produces the best expertise learning. The computational model was used to compare the performance of Drill and Test, Fully Integrated learning, and Incremental Learning. The computational experiments confirmed the limitations of Drill and Test and Fully Integrated learning, and generated the prediction that Incremental Learning produces better performance than the other two instructional methods.

Once the performance predictions were confirmed, the next step is to learn more about the human learning process that takes place during each delivery method. In particular, it is necessary to find out if Incremental Learners develop the meta-cognitive abilities necessary to achieve expertise. This information is important, because, although expertise takes many years to develop, a motivated learner should be able to develop effective problem-solving strategies much more rapidly. Once these strategies are in place, the learner has a greater ability to achieve expertise.

For success in a formal educational environment, an important attribute is the ability to develop study and test-taking strategies effectively under time pressure. The expert interviews and the literature on expertise showed that experts spend their time contemplating problem situations; depending upon the individual, this contemplation appears in many ways, such as comparing and contrasting problems, or analyzing structural features of problems.

Once it is clear that an instructional delivery method encourages effective strategy development, it is reasonable to assume that conceptual development will improve also. So the primary motivation for the human subject experiment is to test the prediction that an Incremental Learning delivery method encourages the development of effective study and test-taking strategies. The experiment also assesses whether conceptual understanding and performance improved more for Incremental Learners than for Drill and Test or Fully Integrated learners.

The human subject experiment consisted of two parts: a laboratory portion in which volunteers took part in a learning task, and a structured interview. Classroom environments and the semester (or quarter) schedule put students in an artificial environment in which they are likely to make strategic decisions based upon habit. Therefore, the learning task had to place the subjects under time pressure as well. The procedures for conducting and analyzing the study were developed using a Meaning Categorization method, developed in this dissertation. This methodology was developed with these requirements in mind. First, Meaning Categorization is a mixed-method interdisciplinary research approach. It can be used to perform both qualitative and quantitative analysis of data: A study can be analyzed from several different perspectives, increasing the validity of the findings. Second, Meaning Categorization is designed to analyze and reveal patterns in qualitative verbal data, such as the interview data in this study. Third, Meaning Categorization can be used to quantify qualitative verbal data, so that statistical analysis can be performed on it. The interview data in this study was quantified and analyzed for statistical significance, to confirm or disconfirm patterns that emerged from the qualitative analysis. In particular, a Meaning Categorization study is an appropriate way to test the prediction that Incremental Learners will develop effective study and test-taking strategies and improved conceptual understanding.

## 5.2 Methodology

An overview of Meaning Categorization is provided first so that the study implementation and results can be understood in their proper context. The study participants, and the experimental procedures which they took part in are then described. The section concludes with a step-by-step description of how a Meaning Categorization analysis was applied to the interview and performance data.

### 5.2.1 Overview of the Meaning Categorization Method

Meaning Categorization is an interdisciplinary method based upon research principles specified by Chi (1997), Kvale (1996) and Miles and Huberman (1994). It combines qualitative and quantitative data collection and analysis. The method is applied primarily to verbal reports, usually interview data, transcribed verbatim from oral or written responses, although numerical data, such as that collected from controlled laboratory studies, can also be analyzed. As with most qualitative research, the number of participants in a Meaning Categorization study is usually small, but supported by a rich volume of data about each participant.

The combination of qualitative and quantitative analysis that Meaning Categorization provides, has several methodological advantages that are consistent with the needs of this study. First, it is a good choice when data is needed about a process as well a product. For example, in order to understand expertise learning, data needs to be gathered about human performance and about human learning. Performance is a product of learning, and often measured quantitatively. Conversely, the process of learning is a complex blend of cognitive and environmental influences, and often assessed qualitatively.

Second, textually rich and descriptive themes emerge from the data analysis. These themes often produce results about phenomena that the researcher did not foresee when designing the study. Research conducted using Meaning Categorization often stimulates many interdisciplinary directions for future study.

Third, Meaning Categorization can test a hypothesis empirically using data from verbal reports. These tests are sometimes formal statistical analysis (such as an ANOVA) and sometimes descriptive (such as averages and standard deviations). Numerical analysis may increase the generalizability of the textual results.

Meaning Categorization should be carried out by at least two researchers, who perform the analysis independently of one another. This approach validates the findings by reducing analytical bias. This method consists of the following steps:

***Step 1: Reduce or sample the verbal protocols.*** Depending upon the amount of transcribed verbal data, select how much of it to code for formal analysis. There are three ways to select data: (1) random sampling of either interviews or portions of interviews; (2) choosing a subset of interviews or portions of interviews based upon some "non-content" criterion, that depends upon the goals of the interview study; (3) perform preliminary coding on the entire content of all the interviews and then perform more detailed coding on a selected subset. This approach is particularly useful for very long interviews. This study used the second method.

***Step 2: Prior to analysis, develop an initial coding scheme that reflects the***

***phenomena being investigated or the hypotheses being tested.*** The coding scheme consists of thematic categories, that correspond to the topics of interest (E.g., conceptual development, strategy development). Decide how to define the size of segments that will be analyzed. Depending upon the goals of the study, the boundaries may be rigid, such as a precise number of words or the full answer to an interview question, or they may be highly flexible, such as when a subject shifts the tone of his or her voice. Steps 1 and 2 are performed together by all the researchers who will be conducting data analysis, guaranteeing that every analyst will evaluate the same data using the same criteria.

***Step 3: Segment the interviews into verbal units of analysis, using the definition in step 2.*** Create preliminary code labels for each verbal unit. Each researcher should conduct Step 3 independently.

***Step 4: Compare the codes to one another in order to identify patterns.*** Develop operational definitions to describe each code. Each researcher should conduct Step 4 independently. Repeat Steps 3 and 4 as often as necessary until all segments are labeled and the definitions are complete and accurate.

***Step 5: Compare the coding decisions and thematic analyses between researchers.*** If they disagree, retain only those codings and themes where both researchers agree. Conduct an inter-rater reliability rating using the following formula:

$$R = A_n/(A_n + D_n)$$

where:

$R$ is Reliability

$A_n$ is Total Number of Agreements

$D_n$ is Total Number of disagreements

Repeat Step 5 until a reliability rating is reached in the range of 90%.

***Step 6: Operationalize the evidence for coding.*** There are many ways to perform this step; which one is the best depends upon the nature of the findings. There is no standard choice of method, although the depiction is often visual. One of the most traditional methods of reporting thematic patterns in qualitative data is through excerpted verbal reports. A select number of unedited participant quotes is presented to illustrate the major thematic conclusions. This approach is powerful, because the data is concrete. Other methods include graphic mental models, cognitive process diagrams, and semantic networks. This study reported thematic data using verbal reports.

***Step 7: Perform a quantitative analysis of the verbal data.*** Use a quantitative analysis to confirm or disconfirm many qualitative findings from steps 3-5. There are many analytical methods to choose from, and which one is the best depends again upon the nature of the data. For example, when three or more sets of data are being compared on some criterion, it is common to conduct an analysis of variance (ANOVA) followed by a post hoc analysis such as the Tukey HSD. These analyses are known as significant difference tests and are performed when there is enough data to satisfy basic requirements for generalizability and statistical power. The results can then be generalized beyond the given sample at hand to the larger population represented by that sample. Because ANOVA, Tukey HSD, and statistical power are fundamental statistical concepts, they will not be explained further in this dissertation. Details of how these methods work can be found in (Glass and Hopkins 1996). ANOVA and Tukey HSD were used in this study.

When there is not enough data to conduct a significant difference test such as ANOVA, perform a descriptive analysis. Descriptive statistics, as implied by their name, are used to describe a set of sample data. Although the results are not generalizable beyond the sample population, they describe characteristics of a sample population that might not have been included otherwise. Sometimes these data reveal emergent trends that do not yet appear in other types of analyses. Other times descriptive statistics give additional context that provides additional insight into qualitative or significance test results. Typical descriptive measures are averages, standard deviations, percentages, medians, and interquartile ranges. All of these descriptive measures were used in this study.

The following sections will describe the participants in the human subject experiment, and how the laboratory learning task and interviews that followed it were conducted.

### 5.2.2 Subjects and Materials

Fifteen volunteers (age 19-51, m = 35 years of age) took part in a one-hour learning study. This number of volunteers was small enough to permit an in-depth investigation of each participant's experiences, yet large enough to allow comparisons between small groups. All of the volunteers were undergraduate or graduate students at the University of Texas at Austin. Participants came from thirteen different academic departments including natural sciences, liberal arts and education. Volunteers responded to advertisements looking for people with an interest in either analytical thinking or mathematics, but who did not know calculus integration. Every volunteer had successfully completed a precalculus course and one semester of calculus which did not include integration. None of the volunteers had mathematics anxiety. Volunteers were not compensated for participating in the study. Each volunteer was randomly assigned to one of three protocols, known to the researcher as Drill and Test (DT), Fully Integrated (FI), Incremental Learning (IL).

Forty-five calculus integration problems were written individually on 4x6 inch index cards. The problems were taken from the set used in the computational study, and were equally divided between three categories (Simple, Usub, Parts). The categories were labeled A, B, C. Alphabetic category labels were used in order to ensure that the participants would not attribute meaning to the category labels. Scrap paper and a pencil were provided.

A set of 4 exams was created for each of the three protocols (12 exams total). The exams contained calculus integration problems that were not part of the study set. The exam problems also were used in the computational study. The fourth exam was identical in each set and consisted of 15 problems equally divided among all three categories of problem. The first three exams varied as follows: for the DT protocol, the first exam contained only Simple Integration problems, the second exam contained only USubstitution problems, the third exam contained only Integration by Parts problems. For the FI protocol, all three of the exams contained a mixture of all categories of problems, equally divided among the three types of problem. The problems were the same on each test, but the presentation order was changed. For the IL protocol, the first test contained Simple integration problems, the second test retained the Simple problems and added USubstitution problems, and the third test retained the Simple and USubstitution problems and added Integration by Parts problems.

### 5.2.3 Experimental Procedure

The first part of the study was a categorization task. Each volunteer was given an identical instruction sheet. The instructions told the participant that they would be given index cards with one Integration problem written on the front of each card, and one of three categories written on the back of each card. Their task was to study the problems and try to identify common properties for each category. The instructions also informed the participant that there would be four timed study periods, each followed by a test; the tests would contain additional problems to categorize.

The study was conducted in a seminar room in order to mimic a familiar learning environment for the students. The length of the study sessions was the same for participants in all three protocols. The first three study sessions were 2, 3, and 3 minutes long. The fourth session was only 1 minute long and simulated a "cram session". These times were determined using pilot studies; the goal was to achieve an optimal balance between applying time pressure, and allowing time to assess the situation and make strategic decisions. Time pressure was desirable for two reasons. First, students enrolled in a class have fixed amounts of time to study prior to taking exams, and need to develop effective ways to use that time. The pilot studies found that short study sessions were most effective at achieving this balance. Second, as discussed earlier, short sessions put pressure on participants to make rapid decisions, react intuitively, and to search for the most effective strategies. When pilot study sessions were long, many participants procrastinated or became distracted by objects in the room or outside the window.

The delivery protocol determined the order in which the calculus problems were presented to the volunteers in each study session. This order paralleled the tests described above. In other words, the DT protocol received one category of problems only in each study session: Simple Integration, then USubstitution, then Integration by Parts. In the cram session they received all three types of problems for study. The FI protocol received all three categories of problems during every study session. The IL protocol received first Simple problems, then Simple and U-Substitution, then Simple, USubstitution and Integration by Parts problems in the third, and again during the cram session.

The second part of the study gathered data via structured interviews. The interviews immediately followed the categorization task. The interviews were developed following guidelines for a thematizing specified by Kvale (1996). Specifically, every participant was asked the same 14 questions about their behavior and experiences (Appendix B). Each question elicited information about one of the key themes of the study: strategy use, reasoning and conceptual development. Many questions asked about the same theme from a different perspective. Follow-up questions were permitted if they clarified previous responses. The interviews were tape-recorded and later transcribed verbatim. Each interview lasted approximately 30 minutes.

### 5.2.4 Experimental Analysis

The steps taken to analyze the interviews using Meaning Categorization were as follows:

***Step (1)*** All of the interviews were selected for coding in order to avoid bias in selection. Responses to demographic questions (Appendix B, questions 7 - 13) were eliminated from coding analysis.

***Step (2)*** An initial coding scheme was developed by two researchers: myself and Mary Z. Last, faculty member at the University of Mary-Hardin Baylor in Belton, Texas. Professor Last has extensive experience analyzing qualitative data. Two thematic code categories, Strategy Development and Conceptual Development, were derived from the prediction being tested by this study: that Incremental Learners will develop more effective study and test-taking strategies, and better conceptual understanding than either Drill and Test or Fully Integrated learners do. Those verbal units that described actions taken by the participants, and that were intended to help them study or take the tests, were coded as Strategy. Coded as Conceptual were those verbal units that described a cognitive state of understanding in regards to their task. A unit of analysis was defined as the amount of verbal utterance used to convey one cognitive ability or behavior. As a result, the size of each unit of analysis could vary from a few words (for example, a participant exclamation) to several sentences (for example, when a participant described a particular study strategy). This grain size makes sense, because the goal of the analysis was to discover each of the cognitive behaviors and abilities used by study participants.

***Step (3)*** Each of the interviews was segmented into verbal units of analysis, by noting every point where there was a change in strategy use or conceptual development. Preliminary code labels were attached to each verbal unit, to identify the behavior or ability. At this stage of analysis the Affective Reactions category emerged.

Table 5.1: **Code List - Interview Analysis.** The list of codes developed from analyzing the interviews. There are three categories of codes: Strategy Development, Conceptual Development, Affective Reactions. Strategy Development codes describe actions taken by the participants that are intended to help them study or take the tests. Conceptual Development codes describe a cognitive state of understanding. The Affective Reactions category captures the unexpected strong emotional reactions that subjects had to the learning task.

| Strategy Development |
|---|
| S-Desire to Solve |
| S-Looking for Rules |
| S-Comparing Group Items |
| S-Reliance on Instinct |
| S-Looking for Patterns |
| S-Reliance on Memory |
| S-Analytical Planning |
| Conceptual Development |
| C-Category Development |
| C-Focus on Complexity |
| C-Lack of Understanding |
| Affective Reactions |
| A-Discomfort |
| A-Positive Feelings |

***Step (4)*** After each interview was fully segmented, the preliminary labels were refined and compared to one another in order to identify thematic patterns. This process was repeated multiple times, until all segments had been labeled. Preliminary operational definitions for each code label were developed and refined at this time.

***Step (5)*** The two researchers met to compare their coding decisions and thematic analyses. Only those codings and themes were retained in which both of us could agree. An inter-rater reliability rating of 90% was achieved. When analysis was complete, there were 12 codes (Table 5.1).

The interview codes are defined as follows:

*Strategy Development*

S-Desire to Solve : The participant explicitly says that s/he wants to, or believes s/he should be able to, calculate or otherwise “solve” the Integration problems in order to categorize them.

S-Looking for Rules : The participant explicitly says s/he is looking for rules that will categorize the Integration problems.

S-Comparing Group Items : The participant is comparing features of individual problems between (inter) or within (intra) category groups.

S-Reliance on Instinct : The participant is hoping for an instinctive or emergent understanding that will guide categorization of the Integration problems.

S-Looking for Patterns : The participant explicitly says s/he is attempting to pattern match in order to categorize the Integration problems.

S-Reliance on Memory : The participant is using a memory-based strategy (primarily memorization) to categorize the Integration problems.

S-Analytical Planning : The participant describes an organized logical/analytical process that s/he is using to categorize the Integration problems.

*Conceptual Development*

C-Category Development : The participant is developing distinct categories for the Integration problems.

C-Focus on Complexity : The participant is using a continuum of complexity to categorize the Integration problems.

C-Lack of Understanding : The participant is making minimal or no progress in learning.

*Affective Reactions*

A-Discomfort : The participant experiences stress and negative reactions to the categorization task. These feelings may range from moderate confusion all the way to high anxiety. They may be implicitly or explicitly stated.

A-Positive Feelings : The participant expresses positive reactions to the categorization task. These feelings may be implicitly or explicitly stated.

***Step (6)*** Verbal excerpts were chosen to represent the thematic results. Participant quotes were chosen because the interviews were very animated; language was the best way to convey their content and energy.

***Step (7a)*** After completing the qualitative analysis, a quantitative analysis of the interview data was conducted. A statistical analysis using ANOVA was performed upon the coded data. ANOVA was selected as a good way to analyze the interviews because data were being compared across three delivery methods, and there were enough pieces of data (verbal units) to confidently generalize the results. The ANOVA revealed where there were statistically significant patterns separating DT, FI and IL protocols. Post hoc Tukey HSD tests isolated the sources of the significant differences.

***Step (7b)*** Descriptive statistics were used to analyze final test performance. Final score distributions were evaluated, and patterns of errors were studied, to look for learning trends. Although the number of participants (15) does not allow for statistical generalizations about performance, a numeric analysis of participant performance does reveal trends that provide additional insight into learning. Median scores, and their associated interquartile ranges (IQR) were computed for final test scores. These two statistics were used (rather than means and standard deviations) because they provide a better descriptor of central tendency when there are a small number of data points, with the possibility that outliers that could mask numerical trends.

The above seven steps describe the procedure used to analyze the human subject experiment. The next two sections will present the results of the qualitative and quantitative analyses developed from Steps 6 and 7.

## 5.3 Qualitative Analysis

This section presents thematic results of the qualitative analysis of participant interviews in each of the delivery methods. Summative descriptions of themes are followed by participant quotes that illustrate them. The qualitative results reveal that Incremental Learning provides a better learning environment than either Drill and Test or Fully Integrated learning.

### 5.3.1 Drill and Test

The Drill and Test participants were extremely nervous. All participants in this group expressed discomfort and anxiety throughout the course of the study. Both their behavior and their language expressed their feelings. For example, two of the students nervously asked the experimenter if she was going to use her masking tape "on them" (the tape was for hanging a Do Not Disturb sign on the door). During the interviews, most students were so anxious that they frequently had trouble expressing themselves:

Student: "and again, I'm, I'm not, I'm a little shaky even on how you, separate them into, these problems, how you separate, what's, you know, where do you put the [making swooping figures with her hands]"

Interviewer: "Parenthesis? That's what you are doing with your fingers there?"

Student: "yeah, yeah"(DT-05)

DT participants lacked organized strategies for studying the categorized problems. Instead, they relied on memory-based strategies, which they were aware were ineffective:

Student: "I never feel like I had really committed the entire category to visual memory...then I was trying to memorize, you know, what the different sets, because within each category it seemed like there were similar cards, sets. So then I was just trying to remember..."(DT-06)

In reflecting during the interviews on the failure of their strategies, three participants were convinced that they must have misinterpreted the instructions, one wondered if she was being tricked, and another complained that the task was unfair. In taking the tests, DT participants took one of two approaches: they either gave up and guessed randomly, or they chose all the same answer on each test. This same answer was the most recently studied concept.

The analysis revealed that DT participants lacked an understanding of how to categorize the integration problems. All DT participants said that they were unsuccessful at learning how to categorize problems; three of them said that they guessed, and had little confidence, on all the tests. One participant acknowledged her lack of understanding as follows:

Student: "I'm aware that my criterion, my criteria are very superficial, and not, I'm, I mean I can tell that they don't work appropriately. Like when I did the test I can tell that it's just, it's not the right criteria." (DT-02)

In summary, all of the members of the Drill and Test group displayed strong negative reactions to the task, relied upon ineffective memory strategies, and developed superficial understandings of the categories.

### 5.3.2 Fully Integrated

The Fully Integrated participants were initially nervous, but their anxiety decreased over time. When the study began, all the subjects were extremely frustrated and overwhelmed. At the end of the first study session, one student burst out into hysterical laughter, one yelled that the task was "impossible! disaster! hopeless!" (FI-01). Another student froze during the first test; she simply sat and waited for the experimenter to return (because the tests were not timed, it was close to 15 minutes before the experimenter went to check on her). Another participant had this reaction:

Student: "[My] impulse, on the first test was to choose all As, because, partly out of frustration...I ended up just going across aesthetically A,B,C,B,A, making a zig-zag." (FI-02)

The analysis of the interviews revealed that three FI subjects gradually evolved a deliberate strategy to look for similarities within groups. Their strategies began to develop sometime after their initial anxiety had partially abated in the second or third study session. The other two FI participants had no specific strategy other than to "just look at them and see if there is anything like a pattern." (FI-05)

All FI participants reported that they mostly guessed on the tests, and predicted that they were not performing well. They did not believe that they understood the categorization task very well. However, four of them were confident that given a lot more time, they could learn to distinguish the categories.

In light of their reported success, four members of the FI group demonstrated an increasing awareness of how the integration problems were categorized. Analysis of the interviews revealed that this understanding was more implicit than explicit. For example, one subject reported that by the end of the study, she was noting regularities on the tests, although she was unsure what to do with this awareness. Another subject showed the beginning of intuitive understanding:

Student: "I noticed...sometimes with the cards, I was having some luck, like I would, after I sort through them all...by the third [study session] I would look at it and I would say "ok I think this is going to be an A: and it WOULD be, you know...but when I looked at the test...I didn't feel confident that I was able to identify those."(FI-03)

In summary, all members of the Fully Integrated group found the task frustrating, by the final study session three of them had developed search strategies, and four of them showed signs of increased understanding of the categories.

### 5.3.3 Incremental Learning

The Incremental Learning group was confident and focused on the task. During both the study sessions and the interview, IL learners made few emotional comments. There was no evidence of fear or anxiety, expressed directly or indirectly through tone or body language. They described the study with words such as "insightful", "fun" (IL-01), "amused" and "stress-free" (IL-03). When pressed by the interviewer, IL participants admitted to being nervous at the start of the study, but reported that these feelings rapidly diminished. The IL group and the DT group shared the same first study session (Simple integration problems only), so it is reasonable that both groups were stressed at first. However, in contrast to the deteriorating attitudes of the DT group, the IL subjects' attitudes improved rapidly. When asked how well they felt they performed on the final exam, four students in the IL group replied with a positive numerical estimate (e.g. 75%). This response is in marked contrast to the DT and FI groups in which all but one participant gave negative verbal estimates (e.g. "pretty bad"). The IL subjects spent the bulk of their interviews confidently describing detailed analytic strategies that they employed to tackle the categorization task. Even when they were not confident that they had succeeded, they were generally confident that they had made solid progress and that given time they would be able to figure out how to categorize the problems. For example:

Student: “The first one [test], I was completely lost...and then [second study session], I was able to compare it and make the correlation...just understand how they were different from one another...the third test, it was insightful, it was a learning experience, I figured out that I still understood A and B...And so I knew, in the final study session...I knew I needed to focus on group C.”(IL-01)

As the above quote demonstrates, study and test-taking strategies in the Incremental Learning group were highly organized and efficient. The participants developed individualized systems that identified subsets of problems on which to focus. They adjusted these subsets in response to new information and insights, or in order to focus on some features about which they were less sure. Two participants systematically moved back and forth between comparing within a group and comparing between groups. They used this process to test and clarify understanding and to reinforce previous conclusions. A third subject devised a system in which she started analyzing the outer edges and general symbols of each problem and moved step by step into the center of the problem and more complex feature combinations. A fourth subject systematically chose two groups at a time to compare, removing from her sight those cards she wished to ignore.

One IL subject differed from the four participants just described, by choosing several successive strategies which relied on memorization and speed. In the interview, this student reported that she knew her strategies were not working. She claimed however, to be noticing some regularities on the tests, although she was unable to explain what she saw.

Analysis of the interviews revealed that the four “successful” Incremental learners were gradually forming a deeper understanding of the integration problems. One student described her progress as follows:

“As we got deeper and deeper into what’s a category B versus a category C, it started getting clear.” (IL-02)

Another subject said that her understanding was “a little better...[then] a little better...”. (IL-03) A third participant felt that she was on the cusp of a breakthrough: “I was looking for [describes features]...but I couldn’t quite find that.” (IL-01)

Another subject gave an example of her categorization when she described how to categorize a sample problem:

Student: “I would definitely put that in a C [Interviewer: why?] Because C was the ones that had e's in them. And, and besides, this is also a more complex of an equation, with both the co-efficient and the exponent. So, the one thing I was noticing about C, was at least to me, Cs had the e's and, and, any, like if you were raising it to a tan, like if the exponent was a tangent or something. Anything that started getting even more complex dealing with e's especially, I would put that in a C. For those reasons." (IL-02)

In summary, all members of the Incremental Learning group had a positive reaction to the categorization task, and all but one of them demonstrated a non-superficial understanding of the categorization. The successful group members employed highly efficient and analytical strategies which reduced the cognitive demands of the task.

## 5.4 Quantitative Analysis

This section presents results of the quantitative analysis of the interview and performance data. Tests of statistical significance confirm the thematic patterns developed by the qualitative analysis. Descriptive statistics analyzed test performance and the results provide additional evidence that an Incremental Learning delivery method is the most effective way to teach for expertise learning.

### 5.4.1 Learning

As expected, subjects in the Drill and Test (DT) and the Fully Integrated (FI) protocols made less cognitive progress than subjects in the Incremental Learning (IL) protocol. Subjects in the IL protocol showed statistically significant differences on several measures of Strategy Development, Conceptual Development and Affective Reactions. The Results are summarized in Table 5.2. A One-Way ANOVA was conducted to examine the differences between the three delivery methods on each of the codes measured in the qualitative analyses. Mean values of analytical planning differed significantly between delivery method ($F(2,12) = 9.33$, $p < .01$). Post hoc Tukey HSD tests indicated that IL subjects had a statistically greater number of analytic strategies than either DT ($p < .01$) or FI subjects ($p < .01$). There was no significant difference between number of analytic strategies used by the DT and FI subjects. These results confirm that learners in an IL learning environment develop better meta-cognitive planning skills than either DT or FI learners.

Mean values of focusing on complexity differed significantly between delivery method ($F(2,12) = 4.56$, $p < .05$). Post hoc Tukey HSD tests indicated that subjects in the IL protocol reported

Table 5.2: **Differences in Strategy Development, Conceptual Development, and Affective Reactions in Drill and Test, Fully Integrated, and Incremental Learning.** Within each thematic category, results are shown for statistically significant differences between delivery method, followed by the source of the difference. For example, under Strategy Development, the first line reports that there is a statistically significant difference in Analytic Planning by participants in the three delivery methods. The probability of this result being in error is $p < .01$. The second line provides additional information about this difference: the IL learners are far more analytical strategy planners than DT or FI learners. The probability of this result being an error is also $p < .01$. The third line reports that there is no statistically significant difference between the analytic strategies used by the DT and FI learners.

| | |
|---|---|
| Strategy Development | |
| Analytic planning differs by delivery method | (F(2,12) = 9.33, $p < .01$)* |
| IL learners more analytic than DT, FI | ($p < .01$)* |
| DT, FI learners equally unanalytic | NSD |
| Conceptual Development | |
| Focus on complexity differs by delivery method | (F(2,12) = 4.56, $p < .05$)* |
| IL learners rely most on complexity analysis | ($p < .05$)* |
| DT, FI learners equally not using complexity | NSD |
| Understanding differs by delivery method | (F(2,12) = 11.03, $p < .002$)* |
| DT, FI learners more confused than IL | ($p < .05$)* |
| Affective Reactions | |
| Discomfort levels differ by delivery method | (F(2,12) = 13.44, $p < .001$)* |
| Positive feelings differ by delivery method | (F(2,12) = 5.57, $p < .01$)* |
| DT, FI learners more uncomfortable than IL | ($p < .01$)* |
| DT, FI learners equally uncomfortable | NSD |
| IL have most positive feelings | (F(2,12) = 5.57, $p < .05$)* |
| * Statistically Significant | NSD: No Significant Difference |

a significantly greater number of conceptual descriptions that relied on complexity analysis than DT (p < .05) and FI users (p < .05). There was no significant difference in the use of complexity between DT and FI subjects. Mean values for lack of understanding differed significantly between delivery method (F(2,12) = 11.03, p < .002). Post hoc Tukey HSD tests indicated that subjects in the DT and FI protocols mentioned significantly far more times that they did not understand the problem than did IL subjects (p < .05) (p < .05). These results confirm that the IL delivery protocol supports cognitive development of complex concepts better than the DT or FI protocol.

Mean values of discomfort differed significantly between delivery method (F(2,12) = 13.44, p < .001). Post hoc Tukey HSD tests indicated that subjects in the DT and FI protocols showed significantly more expressions of discomfort (p < .01) (p < .01) than subjects in the IL protocol. There was no significant difference between expressions of discomfort between DT and FI subjects.

Table 5.3: **Human Performance.** The median final exam scores and interquartile ranges are shown for participants in each delivery method. Drill and Test learners performed the worst, with the lowest median score and very homogeneous results. This result is reasonable, given the well-known failures of Drill and Test instruction. Fully Integrated learners score higher, but vary the most, as is reasonable for immersion learners. Incremental Learners performed the best, with the highest score and a more homogeneous range than Fully Integrated learners. Higher scores and relative homogeneity suggest that the Incremental Learning group shows most promise for continued successful performance.

| | Median Score | IQR |
|---|---|---|
| Incremental Learning | 53.33 | 30.00 |
| Fully Integrated | 46.67 | 36.66 |
| Drill and Test | 40.00 | 19.99 |

These results confirm the results from the qualitative analyses that the DT and FI delivery methods are highly stressful for learners, whereas the IL delivery method is not.

Mean values of positive feelings differed significantly between delivery method ($F(2,12) = 5.57$, $p < .01$). Post hoc Tukey HSD tests indicated that subjects in the IL protocols showed significantly more positive reactions than subjects in the DT ($p < .05$) or FI protocol ($p < .05$). There was no significant difference between expressions of positive feelings between DT and FI subjects. These results confirm the results from the qualitative analyses that the IL delivery method produces a better environment for learning difficult concepts than DT or FI delivery methods.

### 5.4.2 Performance

An analysis of score distribution on the final exam confirmed that IL subjects were making greater cognitive progress than DT or FI subjects. The results are summarized in Table 5.3. Although all of the final scores were low, the median final exam score for IL learners was highest (53.33% compared to 46.67% for FI and 40.00% for DT). Overall the IL learners performed more consistently than FI learners, as reflected in the interquartile range (IQR) of 30.00 for IL learners compared to 36.66 for FI learners. DT learners had not only the lowest median score, but the smallest interquartile range, 19.99, reflecting the homogeneous poor nature of their performance.

Analysis of error patterns made on the final exam further suggests that IL subjects were learning better than DT or FI subjects. There was no discernible pattern to type of error made by the DT subjects. This observation confirms their assertions that they were guessing randomly.

Errors made by FI subjects confirmed their claims that they could identify most of the A category (Simple problems). Most of the errors made by the FI subjects were confusions between the more complex problems: USubstitution and Integration by Parts (categories B and C). However, FI subjects often appeared to be fooled by the length of a problem. They often assumed incorrectly that longer problems had to be more complex. Finally, errors made by IL learners were spread evenly across problem types. IL subjects were somewhat less likely than DT or FI subjects to be fooled by the length of a problem. This finding in particular indicates that the IL learners were beginning to acquire a deeper understanding of the categorization of the problems. The final scores were low because the study sessions were short; however, the IL learners showed a clear trend towards becoming experts. These performance results complement the learning results, and support the hypothesis that IL learners are acquiring the best meta-cognitive skills for learning complex concepts.

## 5.5 Discussion

This study supported the prediction that Incremental Learners develop meta-cognitive abilities necessary to achieve expertise. Incremental learners developed analytic study and test-taking strategies. Drill and Test or Fully Integrated learners developed poor study and test-taking strategies. Incremental Learners also had deeper conceptual understanding of problems than Drill and Test of Fully Integrated learners. Finally, this study established that Incremental Learners had positive affective reactions to learning, whereas Drill and Test and Fully Integrated learners had negative reactions to learning. Whether affective reactions are caused by the delivery method or not is an interesting question for further study. The role of prior knowledge, and the way that conceptual development is reflected in human error, are also interesting areas for further study. These ideas are discussed in this section, along with suggestions for extending the study of Incremental Learning beyond calculus learning.

### 5.5.1 Integrating Qualitative and Quantitative Results

The human subject experiments indicate that meta-cognitive ability is most advanced in Incremental learners. Qualitative analysis revealed that Drill and Test learners rely on their short-term memory, Fully Integrated learners flail and slowly recover, and Incremental Learners develop individualized analytic strategies. Quantitative analysis revealed that there was no statistically significant difference between the strategy development of Drill and Test and Fully Integrated learners, but Incremental Learners were more analytical.

Qualitative analysis found fine distinctions in conceptual development between the three delivery methods. Drill and Test learners both believed and demonstrated that they did not understand the concepts they were trying to learn. Fully Integrated learners occasionally understood some of the concepts implicitly, which contradicted their statements of belief. Their progress is consistent with the observation that by the end of the study some members of this group began to develop organized study and test-taking strategies. As more effective strategies emerged, subtle evidence of understanding began to appear. Quantitative analysis confirmed that, as with strategy development, there were not yet statistically significant differences between the conceptual development of the DT and FI learners. This result explains why FI learners are as yet unaware of having made progress. In contrast, both qualitative and quantitative analysis suggests that conceptual development is improved among IL learners. Qualitative analysis revealed fine details of individual learning strategies, and quantitative analysis confirmed that IL learners were developing a deeper understanding of the complexities of the concepts.

### 5.5.2 Affect and Motivation

The strong emotional reactions experienced by participants in all three groups were unexpected. These results are important for two reasons. First, they provide additional insight into the performance results. For example, the continual anxiety felt by the Drill and Test participants might explain why they relied on memory based study strategies. It may have been impossible for them to think beyond what was right in front of them in each study session. What they saw each time was one group of integration problems that they knew all belonged to the same category. Given that they were under time pressure, they may have instinctively fallen back on a classic study behavior: memorization. Likewise, the Fully Integrated learners' reports of being overwhelmed could make some of their behavior more understandable. Strategies such as choosing test answers so that they create a zig-zag pattern on the answer sheets, or just staring at the problems in the hopes that they would miraculously make sense, reflect a sense of futility. It is hard to imagine any learners developing effective analytic learning habits when they are highly anxious or frustrated.

Conversely, the Incremental Learners' personalized analytic study and test-taking strategies make sense in the context of the enjoyment and lack of stress they reported. Because they rapidly became relaxed, and were challenged by the task, Incremental Learners were able to engage more deeply with the problems. The time constraints did not bother Incremental Learners as much as they bothered the other groups either. This lessened concern about time is additional evidence that they were more internally focussed (i.e. meta-cognitive), a behavior necessary for gaining expertise. A vast body of research has confirmed that deep understanding and long-term learning requires intrinsic motivation. In addition, intrinsic motivation generally co-exists with positive affect. In other words, enjoying study because it is fun or intriguing correlates with positive feelings. Researchers do not yet agree on which comes first, or if in fact they develop simultaneously. However, the most important affective result of this study is that the Incremental Learning delivery method appears to promote positive affect and intrinsic motivation. These attributes are associated with expert behavior, and should be investigated further.

### 5.5.3 Open Questions and Future Directions

There is a lot more to be learned about the conceptual development that takes place using each delivery method. For example, although the participants had similar mathematics backgrounds and attitudes towards mathematics in general, it would be useful to examine what effect, if any, was played by other prior knowledge. One way to examine the role of prior knowledge would be through examining the use of analogy in the interviews. Future interviews could encourage the use of analogy. Another topic to explore further about conceptual development is how it changes during learning in the three delivery methods. There are a variety of ways to obtain data during the learning and test-taking phase of a study; this information will shed greater insight into whether or not learning takes place in spurts of insight or by smoothly increasing understanding, or by some other process. This information is important because it can help educators identify more effective ways to monitor student progress.

Conceptual development can also be viewed by looking closely at the kinds of errors that learners make. This chapter already showed how the seemingly strange test-taking behavior of the Drill and Test and Fully Integrated learners makes more sense in light of their affective states. This information supports other evidence that there was little useful conceptual development in these groups. A closer look at the errors made by the Incremental Learners would increase understanding of what concepts they were developing. In this way, error information can also shed more insight into the performance scores. I predict that additional analysis of the types of mistakes made by Incremental Learners will support the evidence already reported in this dissertation that the IL group is learning better than the other learners. I predict that the IL errors will demonstrate a greater level of sophistication and understanding.

The results of this study should extend to domains other than calculus learning. Therefore, it makes sense to run similar computational and human subject experiments on a learning topic in a closely related field, such as computer science. These results will provide even stronger support for the use of Incremental Learning delivery methods in formal classroom environments. Along the same vein, now that computational simulations and a human subject experiment have supported the use of Incremental Learning, a next research task is to run some experiments in a live classroom environment. Details on how to conduct all of this research, along with other research ideas, will be discussed in Chapter 6.

## 5.6 Conclusions

Evaluation of the human subject results support the claim that neither Drill and Test nor Fully Integrated learning are effective at encouraging the meta-cognitive abilities necessary for becoming experts. On the other hand, an Incremental Learning delivery method works well with time constraints and the pressures that naturally accompany them. The affective results were unexpected, and reinforce these findings. Many new questions about learning for expertise can now be investigated empirically based upon the results of the expert interviews, the computational experiments, and the human subject study reported in this dissertation. The next chapter discusses ways to conduct this research.

# Chapter 6

# Discussion and Future Work

The previous chapters presented several computational and empirical studies in the calculus expertise domain. This chapter begins with a discussion about how the results of these studies support some popular theories of human learning. Then, the chapter overviews future investigations in a variety of areas of human learning, that are opened up by the research presented in this dissertation. First, two ways to analyze the conceptual development observed in the studies. Second, two human subject experiments that will help us understand meta-cognitive development. Third, new computational experiments that extend the computational model into science domains. Fourth, extensions to the computational model that will allow new psychological phenomena, such as the conceptual relationship between integration and differentiation, to be explained.

## 6.1 Relationship to Behaviorism and Constructivism

The research presented in this dissertation has brought together research methods and perspective from computer science, cognitive psychology and education. These results also shed insight into research from educational psychology. Theoretical research into human learning in the 20th century favored first Behaviorism and more recently Constructivism, as the dominant paradigm for evaluating human learning. The results presented in this dissertation correlate with important tenets of these theories. Researchers outside of educational psychology may not know very much about these theories of human learning. One of the high-level goals of this dissertation is to bridge interdisciplinary boundaries and inspire collaborative research. In order to further this goal, it is important to acknowledge the relationship between the dissertation results and Behaviorist and

Constructivist psychological learning theories.

Results from Drill and Test experiments support well accepted understanding about how Behaviorist learning can detract from deep understanding. For example, there are studies of mathematics learning that demonstrate that instructing with this approach produces poor conceptual understanding and encourages memorization (Resnick and Ford 1981; Schoenfeld 1991). In these dissertation experiments, both network and human Drill and Test learners had problems keeping more than one concept at a time in their memory. Reliance on short-term memory strategies resembles some important aspects of Behaviorist psychological learning theory. Classic studies in Operant Conditioning showed that behavior increases in frequency if it is followed by a positive reward (Skinner 1938; Thorndike 1898). When that reward is removed, the behavior rapidly drops to extinction. It is possible to conceptualize what is happening in Drill and Test from this perspective. Each problem set is reinforced for a time, only to be displaced by a new one. There is no motivation for deep learning. The results reported in this dissertation add additional support for ceasing to teach using Drill and Test.

Professional educators will undoubtably agree that it is impractical to teach most academic subjects using full immersion. However, useful perspective on the Fully Integrated results comes from the literature and theory about second and foreign language learning. Although "natural language learning" (as unguided full immersion is referred to in the second language learning literature) can be very effective given enough time, an extensive number of studies show that learning improves even more if some structure is imposed in the early stages of learning (for a review see Spolsky (1989)). The results for Fully Integrated learners reflect this process. Initially, human and network learners experience difficulty with the categorization task. Their performance results are poor. Over time, there are signs of improvement: the network begins to succeed at the categorization task and the human subjects become more comfortable with the task. The network simulations compress time, by simulating long-term immersion. The human learners however, do not have enough time, and so their increased comfort levels do not result in measurable performance improvements. The results reported in this dissertation empirically reinforce the observation that complex academic topics can not be effectively taught using unstructured full immersion.

Results from the Incremental Learning experiments support the Constructivist view that successful deep learning is a self-regulatory process of struggle and conflict. Constructivism grew out of Piaget's well known theory of learning via equilibration (Gruber and Voneche 1995). Existing cognitive concepts have to negotiate with discrepant new insights. The process of negotiation eventually produces new meaning (Fosnot 1996). In this dissertation, the Incremental learners followed more independent and successful learning paths than either Drill and Test or Fully Integrated learners. This was true for both network and human subject learners. All of the human Incremental learners appeared to be more self-aware and meta-cognitive than either the Drill and Test or Fully Integrated learners.

Many educational theorists with a Constructivist orientation also believe in the existence of a "Zone of Proximal Development" (ZOPED) (Vygotsky 1978). The ZOPED is a psychological point on the very edge of a person's current understanding of some concept. If new material is presented below the ZOPED, deep analysis is not required and surface (or no) learning may result. If new material is beyond the ZOPED, the learner is overwhelmed or confused. The implication is that an ideal way to teach is to scaffold the delivery of material at the learner's ZOPED. This will force the learner to struggle with new material in a structured supportive environment. At first the learner will have problems, but eventually she or he will grasp the new concept. Then the ZOPED move upwards and the process can be repeated. The Incremental Learning experiments completed in this dissertation are an example of scaffolded delivery using a ZOPED. Concepts start out easy (Simple integration), and gradually become more complex (Usub, Parts). The learner is forced to compare old concepts with new concepts, but only one new concept is added at at time. The learning task is challenging, but enjoyable, and the subjects make good cognitive progress. The results reported in this dissertation add additional support for the use of Constructivism in the classroom, by showing that an Incremental Learning delivery method has the greatest potential to enhance student learning.

## 6.2 Conceptual Development: Hidden Layer Analysis

This dissertation reported performance and learning results for the ANN computational model. As it learned, the model acquired concepts representing its understanding of how the calculus problems were categorized. This data was stored in the hidden nodes of the network, and changed over time as the network learned. The current results do not include information about what these concepts look like. Right now, we know how well the network learners perform in response to each delivery method, but we do not know a lot about why they performed that way. We need to understand what the concepts look like in order to completely understand learning. Then we can compare this data to the interview data about the concepts learned by humans. This comparison will provide greater insight about conceptual development, that can be used to improve expertise learning.

To analyze the hidden nodes, we need to look at their contents. The hidden nodes contain numeric data, each piece of which represents a small piece of conceptual understanding. These raw data do not make sense to the human eye. However, there are tools that transform the data into a visually meaningful display. The most common method of analyzing the hidden layers of an ANN is Principal Component Analysis (PCA) (Cottrell and Tsung 1989; Elman 1990, 1991a). PCA works by reducing the dimensionality of complex data and extracting features from it. In an ANN, these features correspond to the concepts that the network has learned. Therefore, the conceptual development of the network learning described in this dissertation should first be analyzed using PCA.

Independent Component Analysis (ICA) is a newer, less well-known clustering algorithm (Hyvärinen et al. 2001). ICA is neurally based, and has shown promise in some applications where PCA has failed. Most of the applications of ICA to date have been in engineering (multiple signal discrimination for example) although there have been a few attempts at using it to identify structural patterns in natural language (Bingham et al. 2002; Isbell and Viola 1999; Kolenda et al. 2000). ICA may reveal details of conceptual development missed by PCA. Therefore, the conceptual development of the network learning described in this dissertation should also be analyzed using ICA. Comparing the results of PCA and ICA will provide the greatest amount of information about the conceptual development of the network.

It is important to select the right times for the PCA and ICA analyses. All delivery methods should analyze conceptual development at the beginning and end of training. These data will establish the context for understanding all other conceptual analyses. In addition, the nodes should be analyzed at transition points. These times correspond to important performance checks, and we need to test the prediction that there is a correlation between performance change and conceptual change. If the prediction is confirmed, we will have strong evidence that poor (or good) conceptual development is responsible for the performance results. The additional critical points are as follows:

For Drill and Test, critical points include the transition from Simple to Usub, from Usub to Parts, and the beginning of the cram session. These are the times when test scores changed abruptly. Critical points also include epochs when a performance plateau begins (for those having trouble learning). Examining these locations will allow the comparison of conceptual development between those learners who rapidly acquired the concept and those who did not.

For Fully Integrated, critical points are when the performance begins to plateau, and at one or two points along the plateau. Plateaus need to be sampled to find out whether or not conceptual change occurs even when performance is not visibly changing. I predict that there will be evidence of gradual conceptual change during the performance plateaus. Critical points also occur during the period of rapid increase in performance. Rapidly increasing performance will probably correlate with rapid conceptual change, and this prediction needs to be verified also. Points along a performance plateau or rapid rise can be selected randomly, or they can be selected by examining the performance errors (see next section) for evidence of evolving understanding.

For Incremental Learning, critical points occur at the transition from Simple to Simple-Usub, and from Simple-Usub to Simple-Usub-Parts. As with Drill and Test, these are the times when test scores changed abruptly. Additional critical points include the peaks and valleys of fluctuations during Simple-Usub and Simple-Usub-Parts training. These times may correlate with the acquisition (or loss) of clear concepts. If this prediction is verified, we will have strong evidence that Incremental Learning works by comparing and contrasting difficult concepts until each is clearly identified.

If the predictions made in this section are confirmed, we will know a lot about the relationship between learning and performance that we did not know before. This information will help educators and psychologists understand more about learning for expertise.

## 6.3 Conceptual Development: Error Analysis

This dissertation reported results about the general types of errors and mistakes made by the model and the human subjects. As discussed in Chapter 5, the human Drill and Test and Fully Integrated learners were often fooled by the length of a problem. They also sometimes selected test answers randomly, in response to stress. However, the Incremental Learners selected answers based upon their analytical study and test-taking strategies. We do not yet know if the types of errors made on the tests correlate with changes in conceptual development. This information is important to find out, because errors can tell us a lot more than just how many problems a learner got wrong. They can tell us about the kind of problems a learner is having. Sometimes this information is obvious, e.g. the learner got all the Parts problems wrong, therefore we can safely assume he does not understand the concept of Integration by Parts. On the other hand, some patterns of error are more subtle, e.g. the learner always thought long problems were Parts problems. What does this mean? We will not know until we closely examine the correct and incorrect problems for meaningful patterns.

A good way to map out error patterns is with confusion matrices. This format makes it easy to compare problem types across users and delivery method. I predict that there will be no correlation between error patterns and conceptual development for the Drill and Test or Fully Integrated learners, because they were reacting more to their emotions than to their intellect. However, I predict that there will be a correlation between error patterns and conceptual development for the Incremental Learners, especially on the final two exams.

Confusion matrices will also make it easy to compare the human results to errors made by the ANN. A comparison of human and computational errors will complement the test score and interview data. Together with the hidden layer analysis (previous section), the error analysis will provide greater insight into the learners' performance. When we have analyzed and compared test scores, interview reports, the hidden layers, and the errors, we will have a full understanding of conceptual development for the three delivery methods reported in this dissertation.

## 6.4 Prior Knowledge Through Analogy

Although the subjects selected for the study shared a similar mathematics background and a positive attitude towards mathematics, there may have been other less obvious factors that influenced their learning. Therefore, a next phase of research should look for the influence of more subtle prior knowledge. As discussed in the previous chapter, some of the subjects in the human subject experiment used analogies to describe their understanding and experience. This behavior is expected, because people in general use analogies to relate new concepts to concepts they already understand (Chi and Bassok 1989; LeFevre and Dixon 1986).

A good way to obtain this information naturally and spontaneously, is to encourage subjects to use analogies to explain their strategic decision-making and their conceptual understanding. A greater use of analogy can be triggered by wrapping each integration problem into a word problem. Careful design of the text surrounding each problem can ensure that the contextual cues are balanced across all three delivery methods. This research approach has been used before. For example, studies in analogical transfer of word problem solution strategies have shown that subjects use semantic knowledge to understand and explain a problem's structure. For this example and an extensive review of research in analogical transfer see Bassok (2003). Much of this literature discusses topics similar to the topics of this dissertation, including novice and expert use of deep and surface level structure, and strategy development (Novick 1992).

Adding information about analogy use to the results reported in this dissertation will provide a wide range of new information for improving instructional methods. For example, calculus instructors will be able to explain difficult concepts in ways that will help learners draw upon useful prior content knowledge. They will also be able to design assignments and projects that can be solved using successful strategies students already know.

## 6.5 Affect and Motivation

This dissertation did not set out to study the affective reactions of learners to the three delivery methods. However, as discussed in the last chapter, there were unexpected and clear differences between emotional reactions to the delivery methods. Incremental learners were the only people who developed an intrinsic interest in the categorization task. This is an important result. Several parts of this dissertation have already discussed that experts need to be highly self-motivated learners - this behavior occurs in part because they enjoy what they are studying. Therefore, an important next stage of research is to investigate the relationship between each delivery method, affect, and learning for expertise.

Recent studies of affect distinguish between mood and emotion. Both are subjective states that have an experiential, cognitive and physiological component (Schwarz and Skurnik 2003). Emotion refers to a short-lived reaction to some specific event, whereas mood refers to a feeling with longer duration that is often event-independent. More research has been performed on the relationship between mood and problem-solving than between emotion and problem-solving, and the results vary considerably across problem domains. It is not possible to generalize the relationship between positive and negative mood and learning performance. However, there is evidence that mathematical analysis and logic performance suffers when learners are in poor moods (for a review see Clore et al. (1994)).

Therefore, in designing a study to investigate the relationship between the three delivery methods and affect, two important steps need to be taken. First, using previous research as a guide, the effect of participant mood should be neutralized prior to beginning the categorization task. Doing so will filter out the effects on learning of pre-existing states of mind. Second, interview questions can be designed to directly assess what aspects of the categorization task trigger emotional reactions, and how these emotions shift throughout the study. Obtaining this data will provide greater understanding about how to design the most effective instruction for expertise. It will also contribute to the limited knowledge about the role of emotion and problem-solving. For example, we can gain insight about how much emotion supports or counter-acts instructional methods. If there is an interaction, we can begin to develop ways to use emotional reactions to support successful learning and limit problems.

## 6.6 Computational Experiments in Science Domains

In this dissertation calculus was chosen as the domain partly because mathematics learning applies to many fields of study. The next phase of research should include using the model to simulate categorization of solution strategies for other academic subjects, such as the Natural Sciences.

For example, the first data structures course in computer science would be a logical choice. Drop-out and failure rates in introductory computer science are high, just as they are in calculus. Data structures is traditionally one of the first courses where students are required to think abstractly about complex concepts. This course introduces increasingly complex solution strategies to problems. At some point, the student must acquire the ability to assess a problem and determine the best structure and algorithm to apply.

This learning process has much in common with calculus learning. Human subject experiments can be paired with the computational modeling, in order to provide computer science educators with empirical results to improve educational methods, and student performance.

Chemistry and physics would also be logical choices. These domains use complex mathematical concepts, including calculus. In addition, they begin with fundamental principles of nature, which rapidly increase in complexity. Students branch out into specialties which require them to work with abstract ideas and concrete applications. Chemistry and physics would provide excellent opportunities to evaluate the three instructional methods for expertise learning.

A more ambitious, but exciting area to study would be biology. As a science based upon observations of natural phenomena like chemistry or physics, biology combines complex abstract concepts with real world applications. However, biology includes many concepts with overlapping boundaries (for example, animal classifications) that challenge the learner. Biology would thus make a good domain to study expertise learning in, to extend the applicability of Incremental Learning into less heavily mathematical domains.

## 6.7 Extensions to the ANN Model

### 6.7.1 Generating New Neural Connections

The simulations in this dissertation focused on modeling learning process and performance. Architectural changes can be made to the backpropagation model to capture additional psychological development. For example, cascade correlation networks can simulate the ability of the brain to generate new connections in response to new stimuli. Shultz (1999) used a cascade correlation network to simulate infant habituation to new sentences. In a cascade correlation architecture, the network begins training with a small number of hidden nodes (neurons) and adds additional nodes in response to new stimuli. This initial structure and subsequent growth of the network models the recent recognition that the human brain not only strengthens and weakens neural connections, but adds additional connections throughout its lifespan. Reproducing the simulation results of Drill and Test, Fully Integrated and Incremental Learning with a cascade correlation architecture will open up many new areas for psychological exploration.

### 6.7.2 Linking Integration and Differentiation

One of the higher goals of successful calculus instruction is to enable students to make connections between integration and differentiation. For each integration strategy there is a corresponding differentiation strategy. Historically, mathematics students have been taught a course in integration and a course in differentiation. In many schools, the courses can be taken in either order. The connection between the two classes is left for the student to figure out. If, after having completed both classes, a student can look at an expression and identify underlying mathematical principles linking integration and differentiation, she or he will be able to more easily acquire advanced mathematical concepts that rely upon subtleties of both. It is possible that keeping the classes independent enough to be taken in any order contributes to student failure, in the same way that isolation of concepts in Drill and Test leads to poor learning.

The simulations in this dissertation did not address the issue of relating integration concepts to differentiation concepts. It would be interesting to extend the work to include this type of harder conceptual problem.

Because we know that there is a mathematical relationship between integration and differentiation problems, it is reasonable to hypothesize that there are structural similarities between the problems. To test this hypothesis, and find those similarities, the network can be trained with pairs of integration and differentiation problems that have inverse solution strategies. A Simple integration problem will be paired with a Simple differentiation problem; a Usubstitution problem will be paired with a Chain Rule problem; a Parts problem will be paired with a Product Rule problem.

This task would require the creation of a set of differentiation problems and their corresponding solution strategies. A coding scheme similar to the one for integration problems would have to be developed. The input layer of the ANN would have to be changed to consist of one set of nodes for integration problems and another set of nodes for differentiation problems.

There would be six output nodes in the new network, because the output layer would include both differentiation and integration solution strategies. For each test problem, the correct response would be to identify two answers equally, the integration solution strategy and the differentiation solution strategy that corresponds to it. This is a fundamentally different test procedure, as the desired goal on a test question would be to spread the output activation across two nodes equally, rather than choose one, as in the current experiments.

There is a strong pedagogical reason for studying the relationship between the two calculus topics. Because integration and differentiation classes can be taken in any order, instructors can not assume that students have prior knowledge about the course they are not currently taking. This makes it difficult to know what content to emphasize. However, if simulations reveal structural similarities between integration and differentiation problems, it will be easier to make these decisions.

### 6.7.3 Solving the Integration Problems

The focus of this dissertation was on the first important step in problem solving: selecting the correct solution strategy. The next step for the human learner is to perform the mathematical calculations. Computational models have been developed that simulate calculus computation (Mitchell 1983; Mitchell et al. 1983). These models have often been symbolic rather than connectionist, due to the algorithmic processes inherent in performing calculations.

In future work, it may be possible to train an ANN to step through the process of solving the problem. Simulating this process is important for two reasons. First, just like human learners, the network will learn more about the problem by trying to solve it. It is possible that the preferred solution strategy will change once a solution is attempted. This may happen because applying computational steps to the integration problem may reveal structural features that the network had not discovered during the original strategy selection time. Second, the relative success of a particular solution attempt can be used as feedback for learning to identify solution methods on future problems.

A recurrent network can be used for these simulations. The training process will take place in 4 steps.

1. The ANN chooses a preferred solution strategy, using input and target output similar to those described in this dissertation.

2. The ANN also outputs a series of steps to compute the solution for that strategy.

3. The computation of the solution takes place outside the ANN and the final state of the integration problem is created.

4. The integration problem in its final state is fed back into the ANN as input, along with a flag that reports if the computation succeeded or failed to solve the problem.

Training data can be created using a rule-based system that applies known effective steps to solve integration problems.

By modeling the entire problem-solving process, we will gain insight into learning that extends every aspect of expertise learning addressed by this dissertation. The results of these future experiments will open up many new opportunities for computational and human study.

## 6.8 Conclusions

This chapter discussed some of the promising directions of future research motivated by this dissertation. They continue the interdisciplinary application and methodologies of the completed research. Continuing to pursue both computational and empirical experiments will lead to greater understanding about how appropriate instruction can help people become intellectual experts.

# Chapter 7

# Conclusion

The goal of this dissertation was to understand how instructional delivery methods can help learners acquire the cognitive abilities necessary to become an expert.

Chapter 2 reviewed what we do and do not know about expertise and expertise learning. I discussed early studies of expertise that focused on modeling expert performance. As a result of these studies we know a lot about the cognitive abilities of experts. More recent studies of expertise, in contrast, focus on analyzing and modeling cognitive development. Examples of connectionist models of human mathematics learning were discussed, demonstrating how computational models provide insight into human learning. Human studies of expertise learning need to analyze metacognition and one way to do so is by monitoring strategy development; Recent studies of strategy development were discussed. We now know much about expertise abilities and human development. However, we still do not know enough about how learners acquire expertise, to be able to design effective instructional methods for expertise learning.

Chapter 3 presented a set of results from exploratory interviews with calculus experts. The interviews were conducted using Grounded Theory, a qualitative research methodology that is designed to investigate phenomena without pre-determined hypotheses. In this study, the interviews gathered information about the relationship between how instructors solve problems and how they teach students to solve problems. The interview results led to the development of a theory about how instructional delivery method could affect teaching for expertise. Three hypotheses were developed to test the theory: (1) Traditional sequential delivery methods inhibit learning and retention; (2) Integrated delivery methods increase learning and retention; (3) Incrementally increasing the complexity of the material will lead to the best performance.

Chapter 4 introduced the artificial neural network used to test the three hypotheses. The computational experiments compared learning under three delivery methods i.e. Drill and Test, Fully Integrated, Incremental Learning. Results confirmed the three hypotheses and supported the theory that an Incremental Learning delivery method produces the best learning for expertise. These results then led to the development of a prediction that an Incremental Learning delivery method will encourage the development of meta-cognitive abilities seen in experts.

Chapter 5 presented the methodology of a human subject experiment that tested the prediction. The study evaluated the effect of the three delivery methods on strategy development, conceptual development, and performance. The study was conducted using the Meaning Categorization method. Results confirmed the prediction and also revealed the unexpected result that Incremental Learning produced the most positive affective reactions from learners.

Chapter 6 began with a discussion of how the dissertation results support some popular theories of human learning. Then I discussed the possible interdisciplinary directions for future computational and human subject studies that result from this dissertation. These studies include conceptual development of the ANN, and human studies that investigate other areas of meta-cognitive development such as learning errors and analogy use. The chapter closed with a description of architectural extensions that will permit the model to examine new psychological phenomena such as the relationship between calculus integration and differentiation.

In conclusion, this dissertation presents an interdisciplinary mixed-methodology approach to understanding how intellectual expertise is acquired in a formal setting. The research design drew inspiration from the fields of education, psychology and computer science. Using well established, complementary research methodologies from each field, this dissertation presents a set of results about expertise learning that is not only rigorous, but should be accessible to a variety of researchers. It will therefore provide inspiration and encouragement to other researchers to conduct interdisciplinary research in human cognition.

The results provide insight into how humans learn complex cognitive skills. Qualitative and quantitative analysis of computational and human studies support a computational and empirical theory of learning in which the best learning derives from an incremental delivery strategy. These data suggest how educational delivery methods can help learners develop the meta-cognitive abilities necessary to become intellectual experts. Hopefully, instruction using Incremental Learning will increase the number of students who enjoy their work, pursue advanced studies, and develop sophisticated problem solving skills. Society will benefit from having more intellectual experts ready to address our most challenging problems.

# Appendix A

# Expert Interview Questions

1. What is the typical process that you use for introducing the three methods of integration that we are discussing (Simple Integration, Integration by USubstitution, Integration by Parts)?

2. Why do you introduce integration in the way you just described?

3. What specific difficulties do you see students having with determining the best integration procedure to employ? (as opposed to problems with computing the solution)

4. Is there a particular approach that you teach students to use when approaching any integration problem?

5. (follow on to Q4) Are there any patterns that you tell them to look for?

6. If I were to give you the following 4 examples, how would you approach teaching the students how to decide which method is best to use?

$$\int 2x\sqrt{1+x^2}\mathrm{d}x$$

$$\int_0^{\pi/2} (x + 3\cos(x))\mathrm{d}x$$

$$\int x^2 \cos(3x)\mathrm{d}x$$

$$\int_0^1 e^{x^2}\mathrm{d}x$$

# Appendix B

# Human Subject Experiment Interview Questions

1. How well do you feel you performed on the final test?

2. How much did the final study session help you prepare for the final test?

3. How well do you feel you performed on the earlier tests?

4. Here is a problem similar to the ones you saw on the tests. Please talk me through how you would choose which category it belongs to.

$$\int 3x^2 e^x \mathrm{d}x$$

5. What was your strategy for studying the problems before the tests?

6. What reactions to this task did you have that we haven't spoken about already?

7. What field are you studying in school?

8. What is the Most Recent Math Class that you have Completed?

9. When did you Complete your Most Recent Math Class?

10. Have you completed a Calculus Class?

    10a. If yes, when did you complete your Calculus class?

11. How would you describe your general mathematical ability?

12. How old are you?

13. What ethnic group do you prefer to be affiliated with?

14. Is there anything else you'd like to add before we finish up?

# Bibliography


Abudiab, M. (2001). The impact of technology on teaching an ordinary differential equations course. *JCSC*, 16(3):7–18.

Alvarez, P., and Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: A simple network model. *Proceedings of the National Academy of Sciences, USA*, 91:7041–7045.

Anderson, J. R. (1983). *The Architecture of Cognition.* Cambridge, MA: Harvard University Press.

Anzai, Y. (1991). Learning and use of representations for physics expertise. In Ericsson, K. A., and Smith, J., editors, *Toward a general theory of expertise*, 64–92. Cambridge, UK: Cambridge University Press.

Atkin, J. M. (1998). The OECD study of innovations in science, mathematics and technology education. *Journal of Curriculum Studies*, 30(6):647–660.

Barker, L. J., Garvin-Doxas, K., and Jackson, M. (2002). Defensive climate in the computer science classroom. In *Proceedings of the Thirty-Third Annual Technical Symposium on Computer Science Education*, 43–47. ACM.

Bassok, M. (2003). Analogical transfer in problem solving. In Davidson, J. E., and Sternberg, R. J., editors, *The Psychology of Problem Solving.* Cambridge, UK: Cambridge University Press.

Bednar, J. A., and Miikkulainen, R. (2003). Learning innate face preferences. *Neural Computation*, 15(7):1525–1557.

Bingham, E., Kuusisto, J., and Lagus, K. (2002). ICA and SOM in text document analysis. In *SIGIR.*

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford, UK: Oxford University Press.

Bonnardel, N. (1999). Creativity in design activities: the role of analogies in a constrained cognitive environment. *Creativity and Cognition*, 158–165.

Borko, H., and Livingston, C. (1989). Cognition and improvisation: differences in mathematics instruction by expert and novice teachers. *American Educational Research Journal*, 26:473–498.

Bourlard, H., and Wellekens, C. J. (1990). Links between Markov models and multilayer perceptrons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-12:1167–1178.

Bransford, J. D., Brown, A. L., and Cocking, R. R. (2000). *How People Learn: Brain, Mind, Experience and School.* Washington, DC: N.R. Council National Academy Press.

Bray, N., Reilly, K., Villa, M., and Grupe, L. (1997). Neural network models and mechanisms of strategy development. *Developmental Review*, 17(2):525–566.

Brewer, W. F. (1993). What are concepts? Issues of representation and ontology. In Nakamura, G. V., Medin, D. L., and Taraban, R., editors, *Categorization by humans and machines*, 495–530. San Diego, CA: Academic Press.

Bruner, J. S., and Kenney, H. J. (1965). Representation and mathematics learning. In Morrisett, and Vinsonhaler, editors, *Monographs of the Society for Research in Child Development Ser. 99*, vol. 30-1. Univ. Chicago Press.

Carpenter, T. P., Corbitt, M. K., Kopner, K. S., and Lindquist, M. M. (1980). Results of the second NAEP mathematics assessment: Secondary school. *Mathematics Teacher*, 73(5):329–338.

Chaput, H. H., and Cohen, L. B. (2001). A model of infant causal perception and its development. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, 182–187. Hillsdale, NJ: Erlbaum.

Charness, N. (1991). Expertise in chess: the balance between knowledge and search. In Ericsson, K. A., and Smith, J., editors, *Toward a general theory of expertise*, 64–92. Cambridge, UK: Cambridge University Press.

Chase, W. G., and Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4:55–81.

Chi, M., Feltovich, P., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5:121–152.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *Journal of the Learning Sciences*, 6:271–315.

Chi, M. T. H., and Bassok, M. (1989). Learning from examples via self-explanations. In Resnick, L. B., editor, *Knowing, learning and instruction: Essays in honor of Robert Glaser*, 251–282. Hillsdale, NJ: LEA.

Cimbalo, R. S., and Brink, L. (1982). Aging and the Von Resorff isolation effect in short-term memory. *Journal of General Psychology*, 106:69–76.

Clore, G. L., Schwarz, N., and Conway, M. (1994). Affective causes and consequences of social information processing. In Wyer, R. S., and Srull, T. K., editors, *Handbook of Social Cognition*, vol. 2, 323–418. Hillsdale, NJ: LEA.

Cottrell, G. W., and Tsung, F.-S. (1989). Learning simple arithmetic procedures. In *Proceedings of the 11th Annual Conference of the Cognitive Science Society*, 58–65. Hillsdale, NJ: Erlbaum.

Cromer, A. (1997). *Connected Knowledge: Science, Philosophy and Education*. Oxford, UK: Oxford University Press.

Deci, E. L., Vallerand, R. J., Ryan, L. G., and Richard, M. (1991). Motivation and education: the self-determination perspective. *Educational Psychologist*, 26(3):325–346.

DeGroot, A. D. (1966). Perception and memory versus thought: Some ideas and recent findings. In Kleinmuntz, B., editor, *Problem Solving: Research, Methods and Theory*. Hoboken, NJ: Wiley.

Dörner, D., and Schölkopf, J. (1991). Controlling complex systems; or, Expertise as "grandmother's know-how". In Ericsson, K. A., and Smith, J., editors, *Toward a general theory of expertise*, 218–239. Cambridge, UK: Cambridge University Press.

Duffy, T. M., Lowyck, J., and Jonassen, D. H. (1993). *Designing Environments for Constructive Learning*, vol. 105 of *N.S.A. Division NATO ASI Series F*. Berlin: Springer.

Dyer, M. G. (1983). *In-Depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension*. Cambridge, MA: MIT Press.

Eisenhart, M., Finkel, E., and Marion, S. F. (1996). Creating the conditions for scientific literacy: A re-examination. *American Educational Research Journal*, 33(2):261–295.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Elman, J. L. (1991a). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225.

Elman, J. L. (1991b). Incremental learning, or The importance of starting small. In *Proceedings of the 13th Annual Conference of the Cognitive Science Society*, 443–448. Hillsdale, NJ: Erlbaum.

Erickson, F. (1986). Qualitative methods in research on teaching. In Wittrock, M. C., editor, *Handbook on research in teaching*, 119–161. NY: MacMillan.

Ericsson, K. A. (1998). The scientific study of expert levels of performance: general implications for optimal learning and creativity. *High Ability Studies*, 9(1):75–100.

Ericsson, K. A. (2003). The acquisition of expert performance as problem solving: Construction and modification of mediating mechanisms through deliberate practice. In Davidson, J. E., and Sternberg, R. J., editors, *The Psychology of Problem Solving*, 31–83. Cambridge, UK: Cambridge University Press.

Ericsson, K. A., and Charness, N. (1994). Expert performance its structure and acquisition. *American Psychologist*, 49(8):725–747.

Ericsson, K. A., Krampe, R., and Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100:363–406.

Ericsson, K. A., and Lehman, A. C. (1996). Expert and exceptional performance: evidence on maximal adaptations on task constraints. *Annual Review of Psychology*, 47:273–305.

Ericsson, K. A., and Smith, J. (1991). Prospects and limits of the empirical study of expertise: an introduction. In Ericsson, K. A., and Smith, J., editors, *Toward a general theory of expertise*, 1–38. Cambridge, UK: Cambridge University Press.

Flavell, J. C. (1976). Metacognitive aspects of problem solving. In Resnick, L. B., editor, *The Nature of Intelligence*, 231–236. Hillsdale, NJ: LEA.

Flowers, C. P., Hancock, D. R., and Joyner, R. E. (1991). Effects of instructional strategies and conceptual levels on students' motivation and achievement in a technology course. *Research and Development in Education*, 33(3):187–194.

Fosnot, C. T. (1996). Constructivism: A psychological theory of learning. In Fosnot, C. T., editor, *Constructivism: Theory, Perspectives, and Practice*, 8–33. New York: Teachers College Press.

Gagné, R. M. (1968). Contributions of learning to human development. *Psychological Review*, 75:177–191.

Garner, R. (1990). When children and adults do not use learning strategies. *Review of Educational Research*, 60(4):517–529.

Glaser, B., and Strauss, A. (1967). *The Discovery of Grounded Theory; Strategies for qualitative research*. Chicago, IL: Aldine.

Glass, G. V., and Hopkins, K. D. (1996). *Statistical Methods in Education and Psychology*. Boston, MA: Allyn and Bacon.

Gomez, F., and Miikkulainen, R. (1997). Incremental evolution of complex general behavior. *Adaptive Behavior*, 5:317–342.

Greene, S. (1993). The roles of task in the development of academic thinking through reading and writing in a college history course. *Research in the teaching of English*, 27:46–75.

Gruber, H. E., and Voneche, J. J., editors (1995). *The Essential Piaget*. Aronson.

Hayes, J. R. (1989). *The Complete Problem Solver*. Hillsdale, NJ: LEA.

Hinsley, D. A., Hayes, J. R., and Simon, H. A. (1977). From words to equations: Meaning and representation in algebra word problems. In Just, M. A., and Carpenter, P. A., editors, *Cognitive Processes in Comprehension*, 89–108. Hillsdale, NJ: Erlbaum.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Hoboken, NJ: Wiley.

Isbell, C. L. J., and Viola, P. (1999). Restructuring sparse high dimensional data for effective retrieval. In Kearns, M. S., Solla, S. A., and Cohn, D. A., editors, *Advances in Neural Information Processing Systems 11*, 480–486. Cambridge, MA: MIT Press.

Kolenda, T., Hansen, L. K., and Sigurdsson, S. (2000). Independent components in text. In *Advances in Independent Component Analysis*, 235–256. Berlin: Springer.

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

Kvale, S. (1996). *Interviews: An Introduction to Qualitative Research Inquiry.* CA: Sage.

Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33:1–64.

Lang, S. (1986). *A First Course in Calculus.* Berlin: Springer. Fifth edition.

LeFevre, J., and Dixon, P. (1986). Do written instructions need examples? *Cognition and Instruction*, 3:1–30.

Marshall, R., and Tucker, M. (1992). *Thinking for a Living – Education and the Wealth of Nations.* New York: Basic Books.

Marton, F., and Booth, S. (1997). *Learning and awareness.* Mahwah, NJ: LEA.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457.

McClelland, J. L., Rumelhart, D. E., and the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models.* Cambridge, MA: MIT Press.

Medley, M. D. (2000). Why do they fail? How do they succeed? The success of some students in CS2. In *2nd Annual Computing in Small Colleges Western Conference*, 218–224. JCSC.

Miikkulainen, R. (1997). Natural language processing with subsymbolic neural networks. In Browne, A., editor, *Neural Network Perspectives on Cognition and Adaptive Robotics*, 120–139. Bristol, UK; Philadelphia, PA: Institute of Physics Publishing.

Miles, M. B., and Huberman, A. M. (1994). *Qualitative Data Analysis.* CA: Sage.

Mitchell, T. M. (1983). Learning and problem solving. In *IJCAI*, 1139–1151.

Mitchell, T. M., Utgoff, P. E., and Banerji, R. (1983). Learning by experimentation: Acquiring and refining problem-solving heuristics. In Michalski, R. S., Carbonell, J. G., and Mitchell, T. M., editors, *Machine Learning: An Artificial Intelligence Approach.* San Francisco: Kaufmann.

Moore, T., and Wick, M. (1994). Assessing student's critical thinking skills and attitudes toward computer science. In *The 25th Technical Symposium on Computer Science Education*. ACM.

Murphy, G. L., and Wright, J. C. (1984). Changes in conceptual structure with expertise: Differences between real-world experts and novices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10:144–155.

Nathan, M. J., Koedinger, K. R., and Alibali, M. W. (2001). Expert blind spot: When content knowledge eclipses pedagogical content knowledge. In *Proceedings of the Third International Conference on Cognitive Science*, 644–648. University of Science and Technology of China Press.

Nathan, M. J., and Petrosino, A. (2003). Expert blind spot among preservice teachers. *American Educational Research Journal*, 40(4):905–928.

Newell, A., Shaw, J. C., and Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, 65:151–166.

Newell, A., and Simon, H. A. (1963). GPS: A program that simulates human thought. In Feigenbaum, E. A., and Feldman, J. A., editors, *Computers and Thought*. New York: McGraw-Hill.

Newell, A., and Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nielson, S. H., von Hellens, L., Greenhill, A., and Pringle, R. (1997). Collectivism and connectivity: culture and gender in information technology education. In *Proceedings of the 1997 Annual Conference on Computer Personnel Research*, 9–13. ACM.

Norman, F. A., and Prichard, M. K. (1994). Cognitive obstacles to the learning of calculus: A Krutetskiian perspective. *Research Issues in Undergraduate Mathematics Learning*, 65–77.

Novick, L. R. (1992). The role of expertise in solving arithmetic and algebra word problems by analogy. In Cambell, J. I. D., editor, *The Nature and Origins of Mathematical Skills*, 155–188. New York, NY: Elsevier.

Owen, E., and Sweller, J. (1989). Should problem solving be used as a learning device in mathematics? *JRME*, 20(3):322–328.

Packard, M. G., Williams, C. L., Cahill, L., and McGaugh, J. L. (1996). The anatomy of a memory modulatory system: from periphery to brain. In Spear, N. E., Spear, L. P., and Woodruff, M. L., editors, *Neurobehavioral Plasticity*, 149–184. Mahwah, NJ: LEA.

Paris, S. G., and Winograd, P. (1990). How metacognition can promote academic learning and instruction. In *Dimensions of Thinking and Cognitive Instruction*. Hillsdale, NJ: Erlbaum.

Reed, D. (2000). A perceptual assistant to do sound equalization. In *5th International Conference on Intelligent User Interfaces*, 212–218. ACM.

Reitman, W. (1964). Heuristic decision procedures, open constraints, and the structure of ill-defined problems. In Shelley, W., and Bryan, G. L., editors, *Human Judgements and Optimality*. New York, NY: Wiley.

Resnick, L. B., and Ford, W. W. (1981). *The Psychology of Mathematics for Instruction*. Hillsdale, NJ: Erlbaum.

Robinson, C. S., and Hayes, J. R. (1978). Making inferences about relevance in understanding problems. In Revlin, R., and Mayer, R. E., editors, *Human Reasoning*. Washington, DC: V.H. Winston and Sons.

Rohde, D. L. (1999). A connectionist model of sentence comprehension and production. Dissertation Proposal, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Rosenbloom, P. S., and Newell, A. (1986). The chunking of goal hierarchies: A generalized model of practice. In Michalski, R. S., Carbonell, J. G., and Mitchell, T. M., editors, *Machine learning: An artificial intelligence approach*, vol. 2, 247–288. Los Altos, CA: Morgan Kauffman.

Ross, B. H. (1996). Category representations and the effects of interacting with instances. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1249–1265.

Ross, B. H. (1997). The use of categories affects classification. *Journal of Memory and Language*, 37:240–267.

Ross, B. H., and Spalding, T. L. (1991). Some influences of instance comparisons on concept formation. In Fisher, D. H. J., Pazzani, M. J., and Langley, P., editors, *Concept Formation*. San Francisco: Kaufmann.

Ross, J. A. (1988). Controlling variables: a meta-analysis of training studies. *Review of Educational Research*, 58:405–437.

Rumelhart, D., and McClelland, J. (1987). Learning the past tense of english verbs: Implicit rules or paralell distributed processing. In MacWhinney, B., editor, *Mechanisms of Language Acquisition*. Hillsdale, NJ: LEA.

Rumelhart, D. E. (1998). The architecture of mind: A connectionist approach. In Thagard, P., editor, *Mind Readings*, 207–238. Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In (Rumelhart and McClelland 1986), 318–362.

Rumelhart, D. E., and McClelland, J. L., editors (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA: MIT Press.

Scardamalia, M., and Bereiter, C. (1991). Literate expertise. In Ericsson, K. A., and Smith, J., editors, *Toward a general theory of expertise*, 218–239. Cambridge, UK: Cambridge University Press.

Schoenfeld, A. H. (1991). Mathematics as sense-making: An informal attack on the unfortunate divorce of formal and informal mathematics. In Voss, J. F., Perkins, D. N., and Segal, J. W., editors, *Informal Reasoning and Education*, 311–343. Hillsdale, NJ: LEA.

Schoenfeld, A. H., and Herrmann, D. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, Cognition*, 8:484–494.

Schofield, J. W. (1995). *Computers and Classroom Culture*. Cambridge, UK: Cambridge University Press.

Schunn, C. D., and Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive Science*, 23:337–370.

Schwarz, N., and Skurnik, I. (2003). Feeling and thinking: Implications for problem solving. In Davidson, J. E., and Sternberg, R. J., editors, *The Psychology of Problem Solving*. Cambridge, UK: Cambridge University Press.

Selden, J., Selden, A., and Mason, A. (1994). Even good calculus students can't solve nonroutine problems. *Research Issues in Undergraduate Mathematics Learning*, 19–26.

Shultz, T. R. (1999). Rule learning by habituation can be simulated in neural networks. In *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society*, 665–670. Erlbaum.

Siegler, R. S. (1989). How domain-general and domain-specific knowledge interact to produce strategy choices. *Merrill-Palmer Quarterly*, 35:1–26.

Siegler, R. S., Adolph, K. E., and Lemaire, P. (1996). Strategy choices across the life span. In Reder, L. R., editor, *Implicit Memory and Metacognition*. Erlbaum.

Siegler, R. S., and Shipley, C. (1995). Variation, selection, and cognitive change. In Simon, T., and Halford, G., editors, *Developing Cognitive Competence: New Approaches to Process Modeling*. Erlbaum.

Silver, E. A. (1979). Student perceptions of relatedness among mathematical verbal problems. *JRME*, 10(3):195–210.

Silverman, R. A. (1985). *Calculus with Analytic Geometry*. Englewood Cliffs, New Jersey: Prentice-Hall.

Simon, H. A. (1969). *The Sciences of the Artificial*. Cambridge, MA: MIT Press. First edition.

Skinner, B. F. (1938). *The Behavior of Organisms*. Morgantown, WV: B. F. Skinner Foundation. Reprinted in 1999.

Spolsky, B. (1989). *Conditions for Second Language Learning*. Oxford, UK: Oxford University Press.

Sternberg, R. J. (1998). Metacognition, abilities, and developing expertise: What makes an expert student? *Instructional Science*, 26:127–140.

Stewart, J. (1995). *Calculus*. Brooks Cole. Third edition.

Stigler, J. W., and Hiebert, J. (1999). *The Teaching Gap*. New York: Free Press.

Strauss, A., and Corbin, J. (1998). *Basics of Qualitative Research*. Thousand Oaks, CA: Sage.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review: Monograph Supplements*.

Todd, P. M., and Loy, D. G. (1991). *Music and Connectionism*. Cambridge, MA: MIT Press.

Viscuso, S. R., Anderson, J. A., and Spoehr, K. T. (1989). Representing simple arithmetic in neural networks. In Tiberghien, G., editor, *Advances in Cognitive Science*, vol. 2. Hoboken, NJ and Chichester: Horwood and Wiley.

Vygotsky, L. S. (1978). *Mind in Society*. Cambridge, MA: Harvard University Press.

Warren, J. R., Warren, D. E., and Freedman, R. W. (1993). A knowledge-based patient data acquisition system for primary care medicine. In *Second International Conference on Information and Knowledge Management*, 547–553.

Woodin, D. E. (2001). Design and implementation of GungaWeb: An application of a classical expert system technology to the production of web-based commercial systems. In *The 8th International Conference on Artificial Intelligence and Law*, 104–108.

# Vita

Elizabeth Christine Kaczmarczyk (Lisa) was born on June 17, 1961 in Boston, Massachusetts, daughter of Patricia Harris and Alexander Kaczmarczyk. She received a B.A. in Drama and Spanish from Tufts University (January 1983). Lisa worked in the computer industry before receiving an M.S. in Information Systems from Northeastern University (1990), and an M.S. in Computer Science from the University of Oregon (1992). She spent several years on the Computer Science faculty at Chemeketa Community College in Salem, Oregon. Lisa is an active member of the ACM Special Interest Group on Computer Science Education.

Permanent Address: 11500 Jollyville Road #3624
Austin, Texas 78759
lisak@acm.org
http://www.cs.utexas.edu/ lisak/

This dissertation was typeset with LaTeX 2ε[1] by the author.

[1] LaTeX 2ε is an extension of LaTeX. LaTeX is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The macros used in formatting this dissertation were written by Dinesh Das, Department of Computer Sciences, The University of Texas at Austin, and extended by Bert Kay, James A. Bednar, and Ayman El-Khashab.