NOTE TO USERS

Foundations of Educational Neuroscience: Integrating Theory, Experiment, and Design

Michael W. Connell

Howard Gardner
John Willett
David Rose

A thesis presented to the Faculty
of the Graduate School of Education of Harvard University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Education

2005

For my mother, who gave me Words
For my father, who gave me Ideas
For Rosemarie, who gave me Roots

For Cameron, to whom I bequeath these treasures

Acknowledgements

This thesis represents eight years of work at the Harvard Graduate School of Education. The central questions motivating the work have much deeper roots, however. As a junior high school student, I spent quite a bit of time trying to write simple computer programs on an Atari 400 computer that could carry on intelligent-seeming conversations. My success in that endeavor was minimal, but the exercise raised some very profound questions. For example, if the human mind can be thought of as some kind of "program," how big would the program listing be if it were printed out? What kind of computer could run such a program? What would the programming language (of thought) be like? And the educational question that eventually brought me to HGSE: If we had a programming reference for the mind, would it reveal surprising insights about the mind's workings that we could use to make education more efficient, effective, and individually tailored—or would it just confirm what educators already know and do? My pursuit of these questions during the last two decades has resulted in the manuscript that sits before you. I am thankful to the many, many people—family members, friends, colleagues, teachers, mentors and others too numerous to count—who have contributed directly or indirectly to the final product in one way or another over the years.

A number of individuals and institutions have provided significant direct support during my years of research and writing at HGSE. I am very grateful to Howard Gardner (my advisor), John Willett, and David Rose, the members of my dissertation reading committee, for their patience, generosity, excellent mentoring, and intellectual support throughout the dissertation phase and, indeed, the whole doctoral program. I am also indebted to the Spencer Foundation for their very generous financial support of my doctoral studies in the form of a Spencer Research Training Grant, and to Howard Gardner, Kurt Fischer, Catherine Snow, and Terry Tivnan for supervising my research internships under that grant. The ideas we explored in those internships laid the foundation for this thesis. Catherine Snow also secured the funding that enabled me to attend the Oxford Connectionist Summer School, which was an absolutely pivotal experience in the development of my thinking. In addition, I am grateful to Kurt Fischer for creating many opportunities to present, develop, and refine my ideas through discussions with students and faculty in a constructive and supportive environment.

Mary Helen Immordino-Yang has been a tireless writing partner, and our working sessions have been a critical component of this process. Thanks to Kim Sheridan, Juliana Paré-Blagoev, and Gabrielle Rappolt-Schlichtmann for providing a community and for sharing their time, intelligence, and expertise. I am grateful to Gabbie also for help getting oriented to the statistics software. I am deeply indebted to Theresa Shoemaker for the gift of ten precious days without interruptions to finish writing the first dissertation draft, and to my wife Rosemarie for working to support me during these last seven long months of writing when she would have much preferred to be home with our baby boy. Finally, I would like to thank all of my research subjects, who contributed their time and attention to my empirical study and who in many cases made excellent suggestions concerning refinements to the research paradigm and possible future follow-up studies.

Table of Contents

List of Figures

List of Figures (continued)

List of Figures (continued)

List of Figures (continued)

List of Figures (continued)

List of Tables

Foundations of Educational Neuroscience:
Integrating Theory, Experiment, and Design

Abstract

Neuroscientists and educationists share an interest in learning, suggesting that neuroscience can inform education. Despite educators' eagerness to apply neuroscience to improve their practice, however, few clear examples of such applications exist. Many researchers point to the gap separating microscopic neural processes from macroscopic classroom behaviors as a major obstacle to establishing the neuroscience-education bridge. In this thesis, I describe methods for bridging this gap using computational models to link neural mechanisms to behavioral patterns, and then using these causal neural-behavioral models to inform education. This application of computational models in principle bridges the gap, but in turn raises new issues concerning the validity and interpretability of model properties in relation to human cognition and behavior. The main contribution of this thesis is a set of analytical and experimental tools for addressing the theoretical and practical problems arising in this context.

As I develop the general neuroscience-education tools, I simultaneously demonstrate their application in a specific case. First, I derive a novel analytical framework for comparing disparate psychological theories. I use this framework to justify my selection of an artificial neural network (ANN) over other candidate models, to explain how the ANN relates to human brain and behavior, and to identify a specific neural mechanism that could inform education. Next, I describe an experimental paradigm in which predictions of the neural mechanism are tested against human learning data using multi-level regression models in a novel way to relate ANN behavior to human behavior. Finally, I discuss implications of the research for the educationally relevant

phenomenon of knowledge transfer. The experimental findings are consistent with the model predictions, and on the whole the case study demonstrates the feasibility of using the proposed methods to bridge the neuroscience-education gap in the near term.

Chapter 1

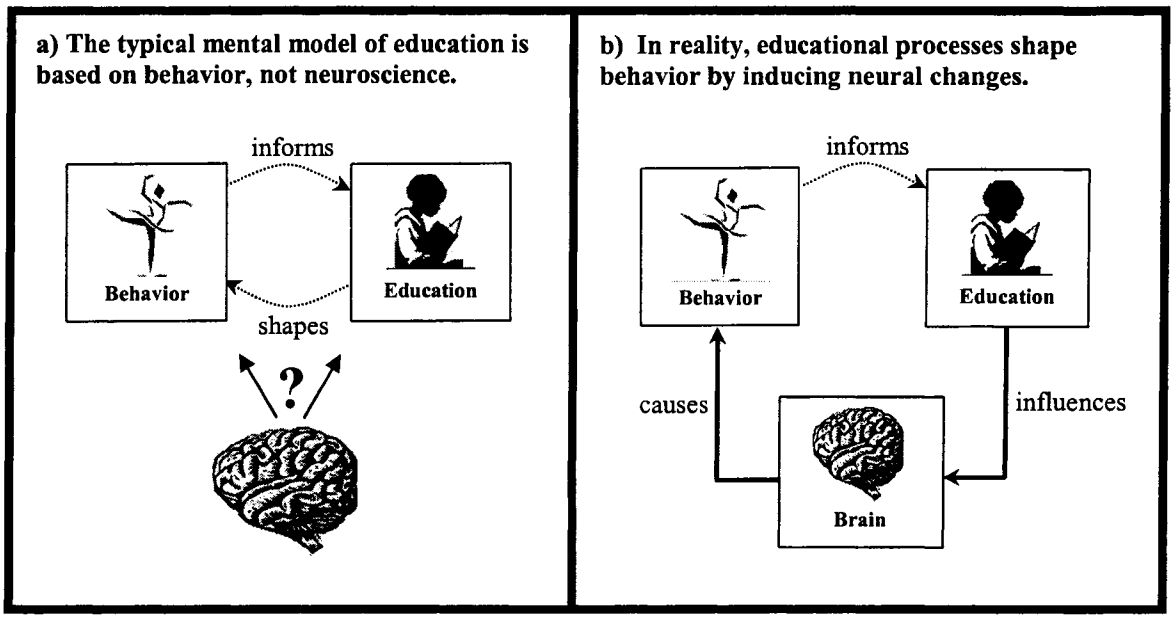
Introduction: Scientific Research in Educational Neuroscience

Background and Motivation

For millennia, mainstream education has been organized almost exclusively around behavioral paradigms. In particular, behavioral data (for example, from classroom assessments or cognitive psychology experiments) provide the basis for educational designs, and interventions based on these designs are evaluated in turn using behavioral outcome measures (for example, gains on a summative assessment). In this view, education is approached as a process of shaping *behavior* toward specific educational goals, using *behavioral* data to track progress and keep students on course toward these goals (Figure 1.1a).

It is true that cognitive psychology has contributed much educationally relevant theory and data over the last fifty years by introducing the "mind" into the discourse and characterizing some of its features (Anderson, 1983, 1995; Chomsky, 1959; Gardner, 1985; Jeffress, 1951; Miller, 1956; Newell & Simon, 1961, 1972). Nonetheless, cognitive psychology's theory of the mind is based on behavioral data (Gazzaniga, Ivry, & Mangun, 2002; Simon, 1992). As such, the construct of the mind in this paradigm is really a systematic and parsimonious re-description of observed behavioral patterns having some predictive power. Cognitive psychology undeniably represents an advance over many competing frameworks for making sense of human behavior—and in particular for informing educational research and design (cf. Bruer, 1993; Carver & Klahr, 2001; McGilly, 1995)—but it is still fundamentally a behavioral paradigm itself.

Figure 1.1: Competing perspectives on brain, behavior, and education



Many educators recognize at some level that the brain must have a role in knowledge acquisition and application. Because the nature of this role is obscure, however, neuroscience does not explicitly inform most people's thinking about education. The problem is that educational interventions do not shape behavior directly; instead, they directly influence unobservable neural mechanisms and processes that in turn generate observable behavior (Figure 1.1b). The brain is unavoidably in the critical path of educational processes, and ignoring it does not nullify its effects.

In the absence of an explicit scientific theory of how neural mechanisms cause observable behavior, educators and educational content producers must be relying on implicit theories of this relationship. We know that intuitive theories tend to be wrong in virtually every domain where they are studied—including physics (Hecht & Bertamini, 2000; McCloskey, Washburn, & Felch, 1983; Reiner, Slotta, Chi, & Resnick, 2000; Viennot, 1979; Zago & Lacquaniti, 2005), biology (Atran, 1995, 1996, 2002; Carey, 1985; Gelman & Raman, 2002; Hamill, 1979; Hatano & Inagaki, 1994; Inagaki & Hatano, 2004; Medin & Atran, 2004), chemistry (Demircioglu, Ozmen, & Ayas, 2004; Galley, 2004; Gopal, Kleinsmidt, Case, & Musonge, 2004; Mulford & Robinson, 2002; Ozmen, 2004), economics (Altmann & Burns, 2005; Brown, 2005; Dunn, Wilson, & Gilbert, 2003; Holzl, Kirchler, & Rodler, 2002; Lerner, Small, & Loewenstein, 2004; Oliver, 2004; Sanford, 2004; Shiv, Loewenstein, & Bechara, 2005), and psychology (Bereiter & Scardamalia, 1996; Clark, 1987; Cosmides & Tooby, 1994; Gilbert, Pinel, Brown, & Wilson, 2000; Gilbert & Wilson, 2000; Goldman, 1993; Haslam, 2005; Kashima, McKintyre, & Clifford, 1998; Malle, 1999; Nichols, 2004; Rosch, 1994). Hence, it would be most surprising if people's spontaneous theories of brain function

were accurate, especially since the human brain is the most complex artifact in the known universe. A major motivation for this dissertation is the idea that education can be improved—perhaps dramatically—if scientifically rigorous insights about brain-behavior relationships can be made accessible to educators.

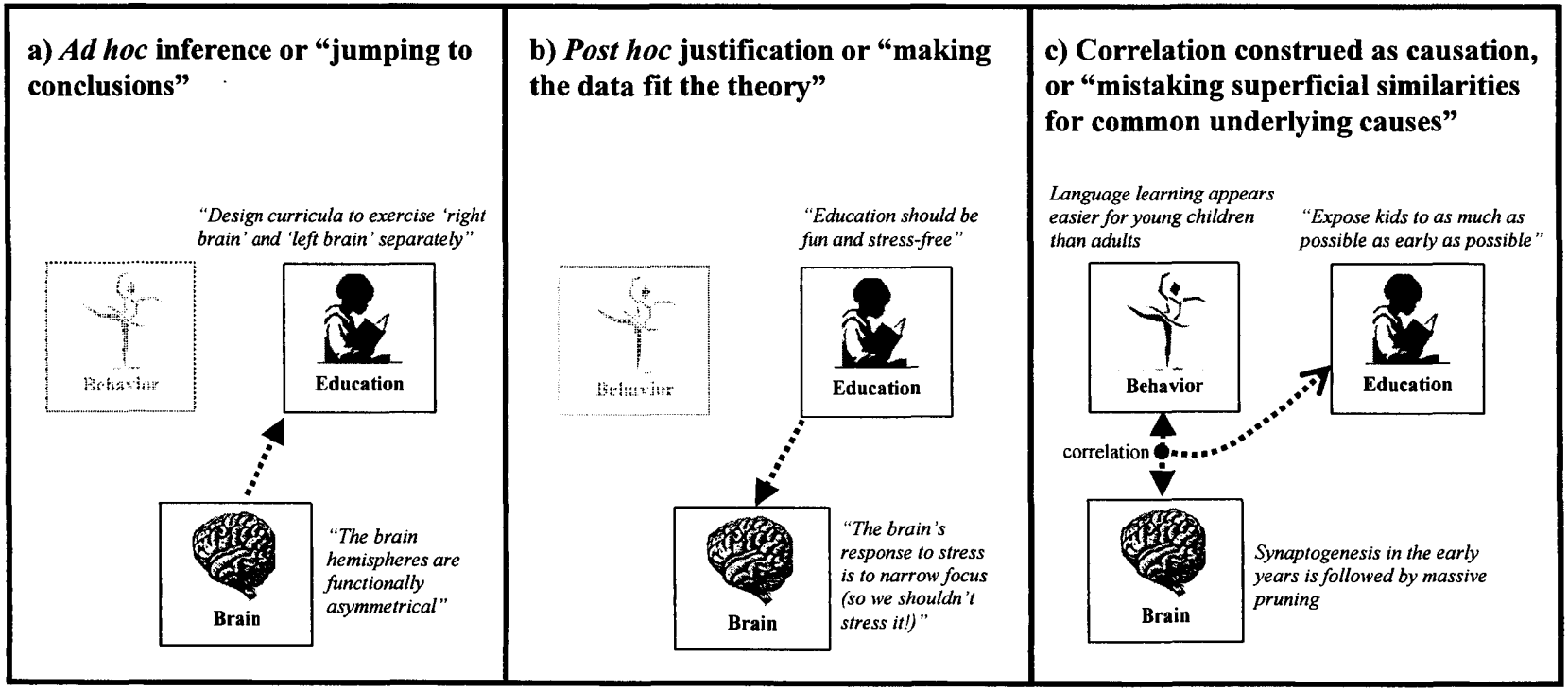
Problems with Current Efforts to Apply Neuroscience to Education

In recent years, people have become enthusiastic about integrating neuroscience into education. In the academic sphere, many universities, schools of education, and research institutes are establishing programs, research agendas, and funding opportunities to support systematic work in this area. Most of this formal research has been initiated fairly recently. Judging by the published educational research literature, these efforts have not yet produced findings that are directly and generally applicable to education (Bruer, 2002), although they should bear such fruit in the coming decades.

In the commercial sector, meanwhile, educational content producers, consultants, journalists, and marketers have identified neuroscience as a powerful tool for exciting public interest and selling products. Unfortunately, this commercial wing of the movement—where claims about neuroscience and education regularly overreach the scientific evidence by a wide margin—is by far the most visible to educators and the general public.

It is instructive to examine the kinds of claims being made in this area to identify common problems and to avoid repeating past mistakes. In my experience, most of the dubious claims being made about educational applications of neuroscience exhibit one or more of the following three common patterns of fallacious reasoning (Figure 1.2): *ad hoc* inference ("jumping to conclusions"), *post hoc* justification ("making the data fit the

Figure 1.2: Three patterns of fallacious reasoning common in the educational neuroscience domain (with examples)



theory"), and/or construing correlation as causation (for example, assuming that similar behavioral patterns must be manifestations of the same underlying neural mechanism).

In cases involving *ad hoc* inference (Figure 1.2a), the neuroscience fact is primary (e.g., "the brain hemispheres are functionally asymmetrical") and the speculative educational implications are meant to follow from it ("e.g., perhaps it would be effective to design separate curricula to exercise the right and left hemispheres individually"). When a neuroscientist or journalist does the speculating, the motivation is often to lend the basic neuroscience research an aura of relevance by trading on the universal and immediate appeal of educational applications. When an educator does the speculating, the motivation is typically to identify novel, "scientifically grounded" educational design principles. The problem is that these speculations amount to raw educational hypotheses that are—at best—a starting point for investigation and that are in any event usually no better than blind trial-and-error search. For example, it is equally plausible to reason from brain asymmetry that perhaps it would be effective to design curricula to foster integration of the asymmetrical functions by exercising them in parallel rather than exercising them individually. The raw fact of asymmetry provides no information about what educational actions might be appropriate. These speculations are starting points for investigation, not final blueprints for educational policy, practice, or design.

Cases of *post hoc* justification (Figure 1.2b) almost always involve an educational researcher, practitioner, or publisher starting with a firmly held value and/or belief about education (e.g., "learning should be fun and stress-free") and trying to validate it by selectively arranging evidence from the scientific literature (e.g., "the brain's response to stress is to narrow its focus") that seems to lead to a prescription in line with the original

belief (e.g., "since the goal of education is to broaden the mind and stress has a narrowing effect, we should therefore seek to make learning fun and stress-free"). The problem is that the evidence can be arranged selectively to tell any story one chooses. For example, starting with the raw fact that stress narrows focus, one could just as easily hypothesize that stress can be used to advantage in educational design. Indeed, the military uses stress very effectively as a central component of its training methods. Similarly, fear-based advertising leverages the narrowing effect of stress to focus consumers' attention on an undesirable possibility and shape their behavior by convincing them that purchasing a particular product will prevent that outcome. While many people might agree that such methods are inappropriate in secondary education, neuroscience offers no support for the view that stress has no educative value (and indeed would probably provide stronger support for the opposite view).

In situations where correlation is construed as causation (Figure 1.2c), a neural mechanism (e.g., synaptogenesis, or the brain's rapid production of synapses in the early years of life, followed by massive synaptic pruning) and a pattern of behavior (e.g., children learn rapidly in the early years of life, whereas the elderly often exhibit mild to severe cognitive decline) are observed to correlate. On this basis a general educational principle or prescription is extrapolated (e.g., "learning is at its peak at the beginning of life and declines thereafter with age, so expose children to as much as possible as early as possible" or, more prosaically, "use it or lose it!").

Such arguments are often reinforced through reference to compelling causal animal models. For example, animal research has established the existence of a critical period in the development of the cat visual system (Wiesel & Hubel, 1965). A cat

deprived of visual input early in life never develops normal vision, even if the animal suffers no permanent physical damage and later gets normal visual input. In the human sphere, it has been observed that young children seem to have greater facility learning second languages than do their adult counterparts (Johnson & Newport, 1989; Lenneberg, 1967). The difference between children and adults is particularly obvious with regard to phonemic awareness and accents, which children seem to acquire readily but many adults learning a second language never fully master. The correlation between a pattern of cat behavior (i.e., a critical period for visual development) and a pattern of human behavior sharing a similar time course (e.g., young children's apparent greater facility with language acquisition, plus a critical period for readily acquiring accents) has led people to conclude that the same neural mechanism *causes* both sets of behavioral results. From there it is a short hop to propose that children should be exposed to foreign languages (or any content, for that matter) as early as possible, or risk losing the ability to fully master them forever.

This argument might seem, on the face of it, to employ more sound reasoning than in the previous two cases. The weighty inferences from cat physiology to human physiology and from a sensory function (visual processing) to a higher cognitive function (language) are based, however, entirely on superficial correlations between visual behavior in cats and linguistic behavior in people. Despite appearances, the critical inference in this argument is not well supported and is quite dubious. Indeed, some evidence suggests that adults can learn second languages as easily as or more easily than children can when one controls for the immersive learning environment in which children typically learn language (Abu-Rabia & Kehat, 2004; Marinova-Todd, Marshall, & Snow,

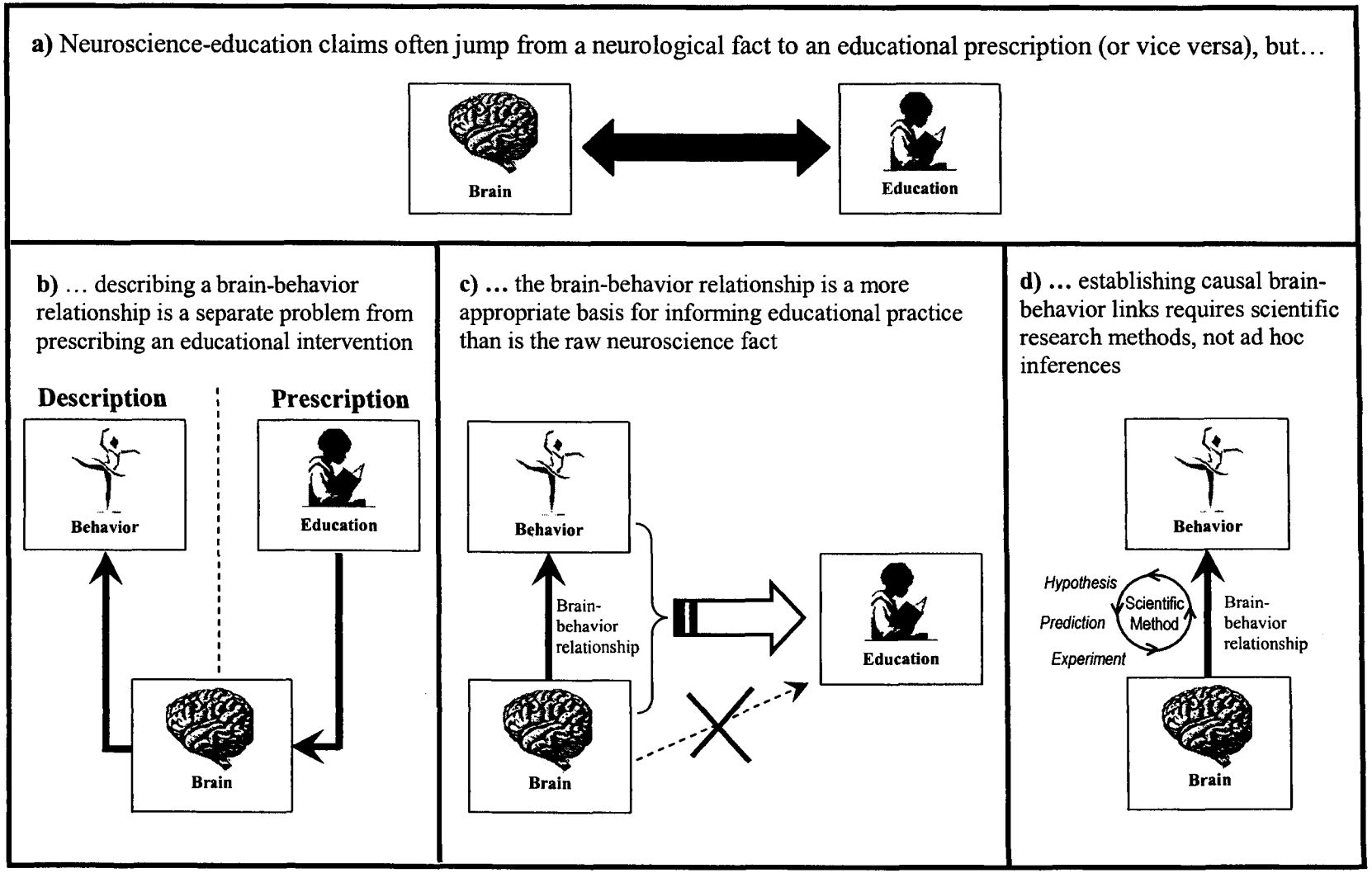
2000, 2001; Snow, 1983, 1992, 2002; Snow & Hoefnagel-Hohle, 1977, 1978a, 1978b, 1979). This case illustrates why rigorous methods and standards are needed in applying theory to educational design and practice. Only in this way can we avoid costly decisions based on enticing but ill-supported notions that later turn out to be wrong—whether informed by neuroscience, cognitive psychology, intuition, or any other source.

Lessons Learned: General Criteria for a Rigorous Educational Neuroscience Framework

The negative examples just reviewed illuminate issues that must either be avoided or faced head-on by a rigorous neuroscience-education research framework. One general insight is that people get into trouble when they make the mistake of attempting to jump directly from neuroscience facts to educational prescriptions or vice versa (Figure 1.3a). In my view, there are at least three major problems associated with this practice that must be avoided.

The first problem is that direct inference from a neuroscience fact to an educational application conflates two qualitatively different steps: description and prescription (Figure 1.3b). Theories of how a system such as the brain operates internally (basic descriptive or explanatory theories) are logically distinct from theories specifying how the operation of that system might be manipulated through external means (applied theories of intervention). In other words, understanding how the brain produces a specific behavior is necessary but not sufficient for determining how specific educational interventions will interact with the brain to systematically change that behavior. In other words, the entire neuroscience-education circuit involves two sets of causal relationships: a) the relationships between specific neural mechanisms and the observable behaviors

Figure 1.3: Three problems with attempts to jump directly from neuroscience facts to educational prescriptions



they produce, and b) the relationships between specific educational actions (designs, methods, interventions, etc.) and the neurological changes they induce—which ultimately result in changed behavior. Despite the efforts of many people who try to collapse these two considerations into one step (recall, for example, Figures 1.2a and 1.2b), these two steps require independent investigation and systematic validation.

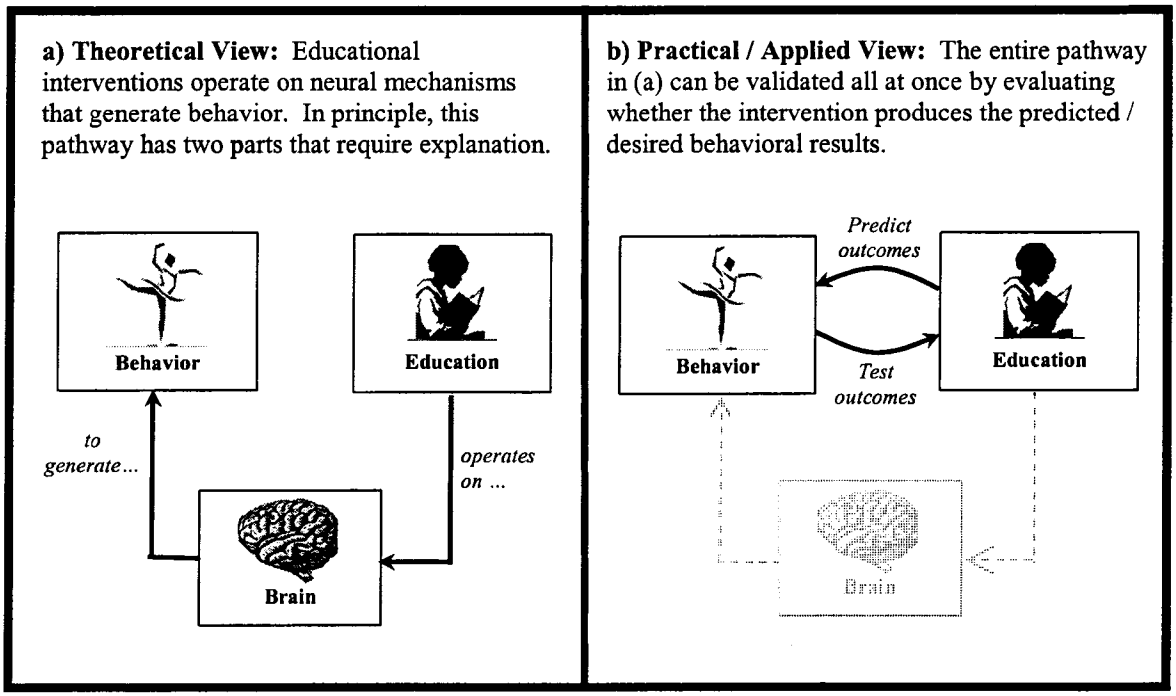
The second problem becomes apparent once the brain-behavior relationship is distinguished from the educational intervention. The problem stems from the assumption that educational prescriptions can be derived from raw neuroscience facts, even though bald facts about the brain (for example, the fact of hemispheric functional asymmetry) generally do not provide information about appropriate educational actions (for instance, whether to exercise the hemispheres separately or together). Without a model of the behavioral changes that would result in each case, the brain data themselves are mute regarding an appropriate course of educational action. For this reason, I would argue that a more appropriate basis for educational applications is the *relationship* between a neural mechanism and the pattern of behavior it causes, because this relationship provides much more insight concerning the educational actions that are likely to produce desired behavioral effects (Figure 1.3c).

The third problem pertains to the validity and credibility of a given claim about neuroscience and education. It is not enough simply to *state* that neuroscience supports a particular educational design. Introducing neuroscience into the process of educational design is entirely gratuitous unless it can be *shown* how the behavioral implications follow systematically from the neuroscience facts. Drawing *ad hoc* inferences about behavior directly from facts about the brain basically circumvents the entire chain of

reasoning that would constitute a sound and valid scientific argument. The serious investigation of causal brain-behavior links that could provide a grounded basis for educational practice will require much more rigorous methods than those used in the examples just described (Figure 1.3d).

Ultimately, it would be useful to conduct basic research to investigate how educational interventions interact with neural mechanisms to do their work (that is, to produce a detailed theory of the full pathway in Figure 1.4a). In the meantime, however, a more expedient approach to validation should be possible when the educational goal is practical (e.g., to produce specific behavioral educational outcomes) rather than theoretical. Once a neural-behavioral mechanism has been identified and used to inform educational designs, in general it should be possible to evaluate the resulting educational designs using well-established, purely behavioral methods. In effect, the entire pathway from educational intervention through neural mechanisms to observable behavior (Figure 1.4a) is evaluated all at once by testing the predicted behavioral outcomes produced at the final step (Figure 1.4b). If an intervention produces the expected behavioral outcome, then the applied educational goal is achieved, and further research at the neural level is unnecessary. If the expected effects are absent, then further investigation of neural mechanisms might be warranted to refine the neural-behavioral theory in order to improve the educational design. The behavior-level evaluation of outcomes is likely to be faster and more straightforward than building a detailed neural theory in any particular case. As such, behavioral evaluation methods provide a critical component of an efficient strategy for moving toward rigorously generated usable knowledge in

Figure 1.4: Theoretical vs. applied strategies for validating educational interventions



neuroscience and education relatively quickly, without having to wait for scientists to work out every theoretical detail of the physical processes involved.

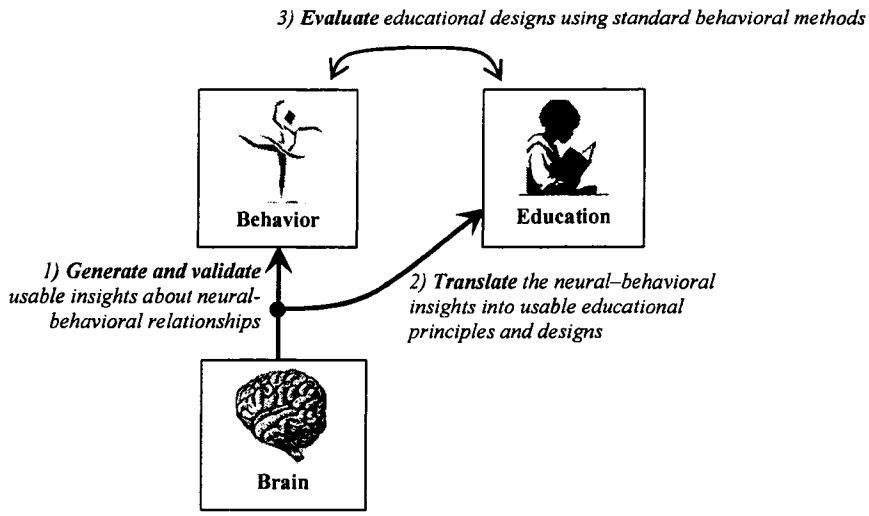
Proposal for a Minimal Educational Neuroscience Research Framework

Based on the preceding analysis, I submit the framework depicted in Figure 1.5 as a minimal framework for conducting rigorous scientific research in the domain of neuroscience and education (hereafter referred to as "educational neuroscience"). In this paradigm, the process of applying neuroscience to education involves three major steps:

1) Characterize causal brain-behavior relationships
2) Identify educationally relevant implications from step #1 and use these to inform the design of educational materials, methods, experiences, and/or environments
3) Validate the designs in step #2 using standard behavioral methods (e.g., experimentally controlled outcome studies)

The third step in this framework (evaluating educational designs) involves well-established behavioral methods; accordingly, I do not discuss it further in this dissertation. The particular educational principles that can be extracted in the second step depend on the details of the brain-behavior link(s) identified in the first step in any specific case. I demonstrate how this can be done using a concrete case study. My proposal for identifying, characterizing, and validating potentially useful brain-behavior relationships (step #1) is the most novel and by far the most challenging part of this process, so I focus in this dissertation primarily on identifying and addressing the obstacles that arise in connection with executing it. In the next section, I describe this part of the educational neuroscience framework in more detail.

Figure 1.5: A minimal framework for conducting rigorous applied research in educational neuroscience



A Scientific Method for Basic Research on Brain-Behavior Relationships

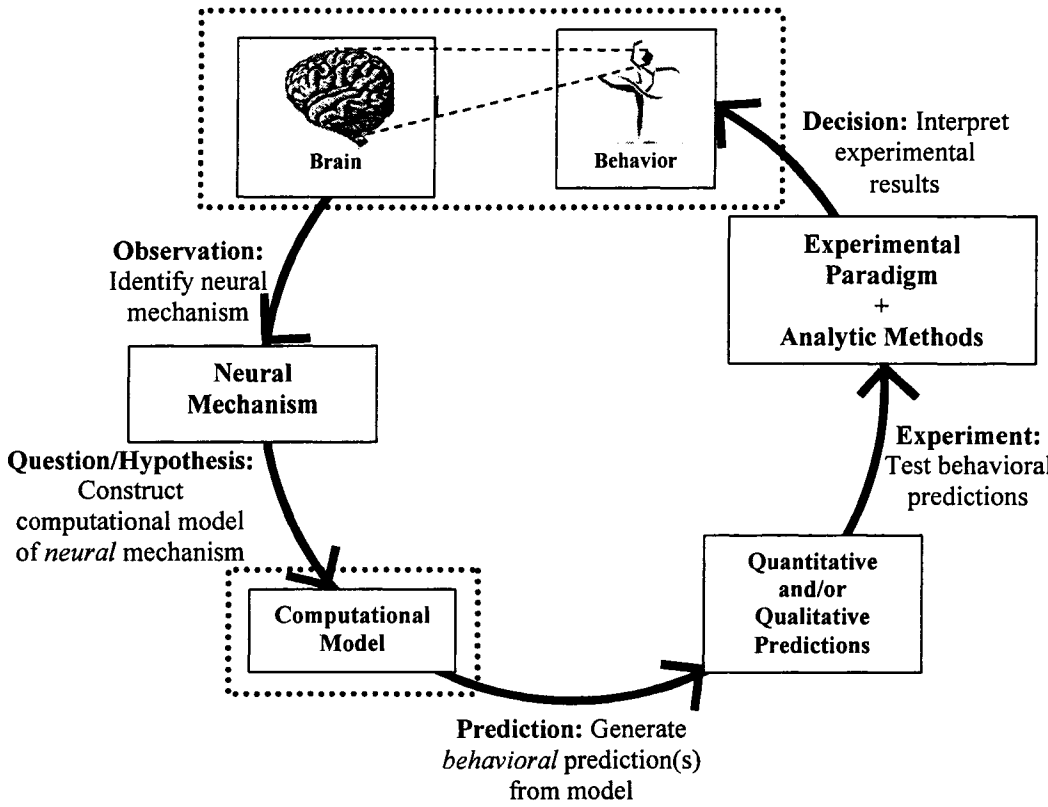
Scientific research in any domain typically involves a number of steps: observation of a phenomenon (e.g., bricks fall faster than feathers), generating a question ("does the weight of an object determine how fast it will fall?"), hypothesizing an answer to the question ("an object falls at a rate proportional to its weight"), predicting the outcome of an experiment based on the hypothesis ("if I drop a two ounce weight, a one ounce weight, and a one ounce feather from the same height, the larger weight should hit the ground in half the time of the smaller weight and the feather, which should land together"), conducting the experiment (drop the three objects from a high window and observe that the two weights land together and the feather lands much later), and a decision about the likely veracity of the hypothesis based on the evidence (in this case, the evidence would be inconsistent with the hypothesis and the hypothesis would probably be rejected, or at least deemed highly unlikely).

A special challenge inherent in any attempt to identify causal brain-behavior links is the problem of bridging the very different temporal, spatial, and organization scales separating neurons and synapses on the one hand from complex patterns of behavior on the other. I argue that this can be accomplished through the use of computational models (such as artificial neural networks or dynamical systems models) that can embody neural mechanisms and generate behavior-level patterns of activity. Specifically, I propose the method outlined in Table 1.1, adapted from the general scientific method, for conducting research on causal brain-behavior relationships (see also Figure 1.6).

Table 1.1: A scientific method for conducting brain-behavior research using computational models

Observation: Identify a potentially interesting neural property, mechanism, etc.
Model: Specify a computational model (such as an artificial neural network, a dynamical systems model, or a production system) that embodies the crucial features of the neural property or mechanism under consideration. Also identify one or more patterns of model behavior that are caused by the modeled neural mechanism.
Question: Specify a question linking the neurological properties to behavior or linking properties of the model to characteristics of people. For example: • Is the type of mechanism represented in the model actually present in the human nervous system? • Does the neural mechanism have the same behavioral consequences in people that it does in the model?
Hypothesis: Depending on the observation and question in a particular case, specify an appropriate hypothesis. For instance: • The same mechanism represented in the model is present in the human nervous system. • The specified neural mechanism causes patterns of behavior in people analogous to the behavioral patterns observed in the model.
Prediction: Use the computational model to generate testable behavioral predictions about human learning and cognition that are causally related to the specific neural property or mechanism under study.
NB: The computational model must be handled carefully in this framework, because there are many model properties and behaviors that are merely artifacts of the model and thus have no bearing on human cognition or behavior. For the purposes of this framework, model properties and behaviors can be separated into two categories. The first category is comprised of all the model properties that are identified with the neural mechanism under study and all the model behaviors that it causes. The second category contains everything else, including potentially valid predictions of the model that are not being studied at the time as well as model artifacts, etc.
Experiment: Test the behavioral predictions using empirical behavioral data from people.
Decision: Decide whether to reject or revise the hypothesis based on whether the experimental evidence from people is consistent with the model predictions or not.

Figure 1.6: A general framework for conducting basic brain-behavior research, based on the scientific method (step #1 in Figure 1.5)



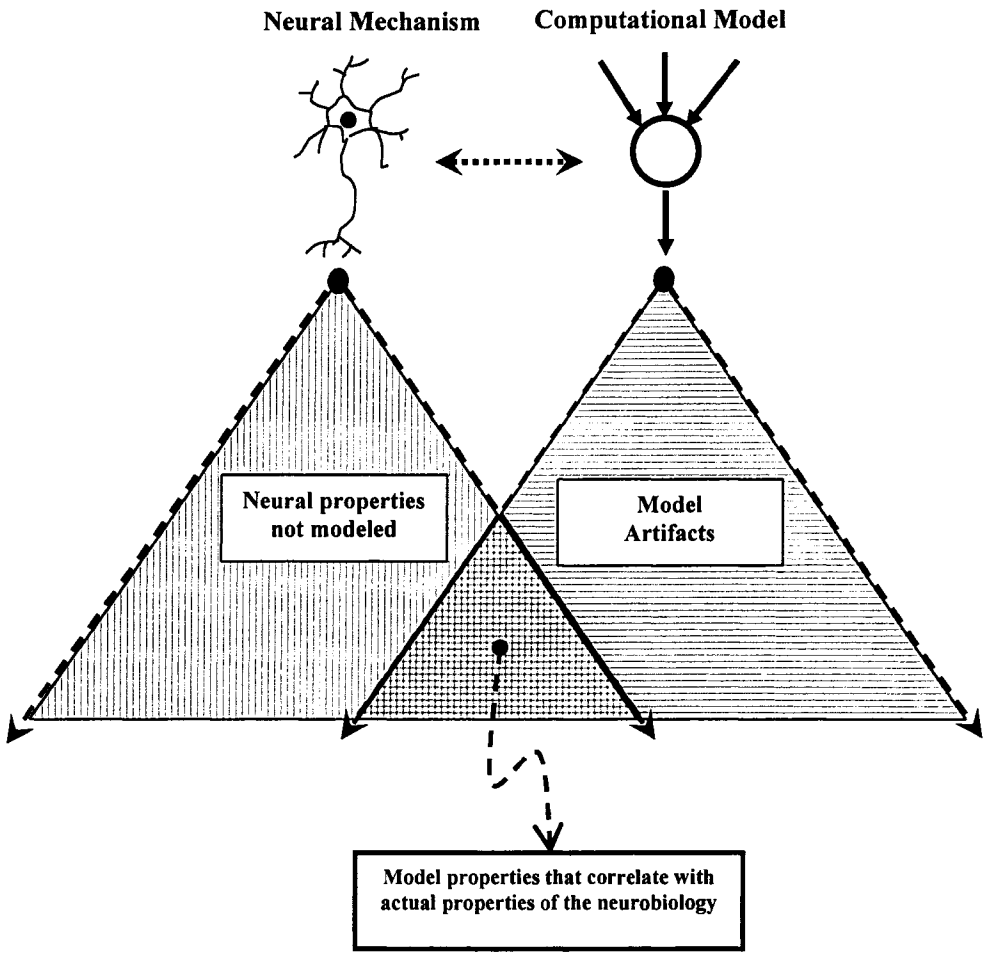
The computational model is crucial in this framework for linking neural mechanisms to behavioral implications. My suggestion is that the computational model should be constructed so that it embodies the neural mechanism identified in the observation step, but also in such a way that it can generate behavior-level predictions that follow from the embedded neural mechanism. The brain-behavior link is established in this manner. Rigor is maximized by considering only those model properties that correspond with the neural properties under investigation and on model behaviors that they cause (Figure 1.7). Model validity is tested formally using quantitative methods and empirical data.

The validation strategy outlined in Table 1.1 is based on the scientific principle of *falsification*. Ideally, the falsification step is executed using quantitative methods to conduct formal hypothesis tests of *a priori* behavioral predictions, and the results of such tests have implications for the entire modeling framework instead of just the specific model being tested. This feature distinguishes the current approach from most other applications of computational models. Typically, these models are used to provide proofs-by-example or other *positive* demonstrations that a particular computational modeling framework is sufficiently powerful to generate a specific set of data.

Using computational models as a tool for investigating brain-behavior links in this manner in principle addresses the problem of establishing causal relationships across the very different levels of analysis involved, but this proposal raises three new questions:

1) What is the theoretical status of computational models (for example, ANNs)? That is, how should we understand the models as theories of neural and psychological function and observable behavior?

Figure 1.7: Computational models require careful interpretation. The neural mechanism and the computational model of it each have a "cone of implications." Some properties of the neurology are not included in the computational model ("neural properties not modeled"), and some model properties are artifacts of the model instead of implications of the neural mechanism embedded in it ("model artifacts"). The overlap of these two cones defines the space of valid possibilities for making inferences from model behavior back to the biological system (cross-hatched area where the two cones overlap).



2) How can we identify model properties and behaviors that represent a basis for valid inferences to humans (given model artifacts, model incompleteness, etc.)?
3) How can we verify the models against empirical data on human subjects?

My efforts to answers these questions constitute the bulk of work in this dissertation. In part, this is because to my knowledge none of them has previously been answered in a way that provides both an analytic or theoretical justification and a blueprint for operationalizing them in practice (e.g., for identifying educationally relevant neural mechanisms, designing experiments, drawing out educational principles, etc.). I therefore had to develop new analytic and experimental tools (and/or adapt existing ones to this novel domain) suitable for addressing these questions.

Applying the Brain-Behavior Research Method: A Case Study

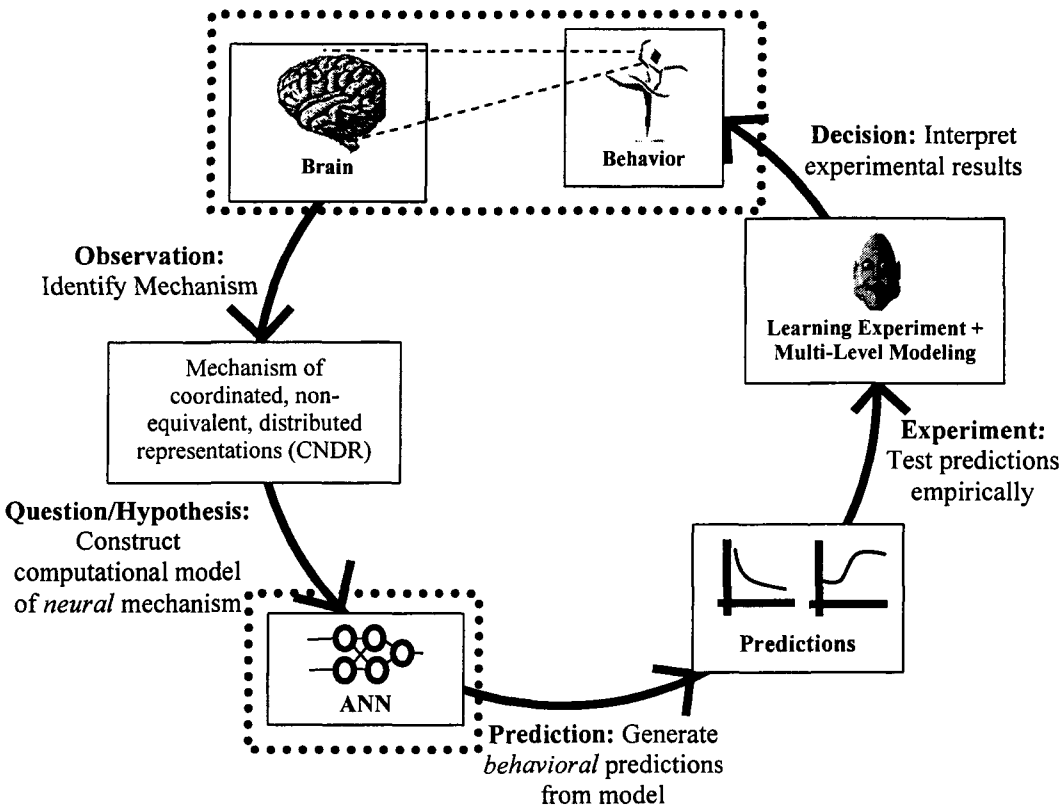
The general research method described in the previous section can be particularized in many different ways (for example, using a variety of different computational models, generating a range of predictions, using a number of appropriate experimental paradigms and analytic tools, etc.). In addition to developing a set of general tools, in this dissertation I also work through a concrete case study to demonstrate one way that this entire process can be carried out (see Table 1.2 and Figure 1.8).

In parallel with this basic research process (since many such experiments will generally be necessary to establish a particular brain-behavior relationship), educational implications of the neural mechanism and its behavioral consequences can be deduced and applied to educational designs, which can then be evaluated independently using standard behavioral methods.

Table 1.2: A case study demonstrating the application of the brain-behavior research framework

Observation: Drawing on findings from wet neuroscience, I first isolate a mechanistic principle that seems to be fundamental to the way biological neural networks represent and process information. That is, these systems employ two different types of distributed representations (synaptic weights and patterns of spreading activation) that are not copies of one another but contain different information (as distinguished from representing the same information in different formats). The special way these two sets of non-equivalent distributed representations are coordinated in a neural network is what underlies many of their interesting properties (for example, graceful degradation, spontaneous generalization, and content-addressable memory). I refer to this as the mechanism of "Coordinated, Non-equivalent, Distributed Representations" (or CNDR for short).
Model: Many (if not all) artificial neural network models (ANNs) already embody the CNDR mechanism, so in this case I can simply select one of these rather than constructing my own. I choose the connectionist model (or multi-layer perceptron) since it has been extensively studied and is often used to model behavioral tasks.
Question: Does the human nervous system employ a CNDR mechanism like the one represented in a multi-layer perceptron (or connectionist ANN)?
Hypothesis: The human brain uses a CNDR mechanism with the same general properties as that represented in the ANN for learning novel categories.
Predictions: I designed a category learning task and trained the ANN on it. Based on an analysis of the ANN behavior during and after learning, I generated two quantitative behavior-level predictions: a. People's reaction times should decrease curvilinearly with stimulus distance from the category boundary (in stimulus feature space). b. As learning progresses, people's perceptual judgments of similarity should change systematically as a function of the categories being learned.
Experiment: I designed an analogous version of the category learning task to administer to human subjects in an experimental setting. I identified multi-level regression modeling as a statistical method that could be used to test the two ANN predictions quantitatively using empirical data on human learning.
Decision: The experimental findings are consistent with the hypothesis that human learning and internal organization of novel categories rely on a CNDR neural mechanism like the one represented in the connectionist ANN.

Figure 1.8: Example of an application of the proposed method to investigate a specific causal brain-behavior relationship



Organization of the Dissertation

The rest of this dissertation is organized as a set of four papers describing general tools for addressing key problems in this domain and demonstrating how those tools and the research framework described in this introduction can be applied in a concrete case of educational neuroscience research.

In the first paper (Chapter 2), I develop an analytic framework enabling the uniform comparison of disparate psychological models, theories, and paradigms. I apply this framework to compare and contrast four psychological and behavioral models: 1) the behaviorist stimulus-response model, 2) the perceptron neural network, 3) the cognitivist production system (as an exemplar of the symbolic paradigm), and 4) the multi-layer perceptron (MLP) neural network. Through this analysis, I identify a specific neural mechanism worthy of study (the mechanism of coordinated, non-equivalent, distributed representations, or CNDR for short) and justify my use of a multi-layer perceptron to generate behavioral predictions from the CNDR neural mechanism.

In the second paper (Chapter 3), I describe behavior-level patterns resulting from the CNDR mechanism as embodied in an MLP neural network model, and I propose an experimental paradigm and an application of multi-level regression modeling to test the validity of this hypothesized brain-behavior relationship in people. I also report on the results of conducting the experiment. Together, the first two papers elaborate on, justify, and demonstrate the application of the general brain-behavior research framework described in the previous section (which constitutes "step 1" in the educational neuroscience framework of Figure 1.5).

In the third and fourth papers, I discuss potential educational implications of the findings from the first two papers ("step 2" in the neural-educational research framework of Figure 1.5). In particular, in the third paper (Chapter 4) I argue that different assumptions about the brain lead to conflicting behavior-level models of educationally relevant phenomena such as knowledge transfer. In other words, not all possible theories of transfer are compatible with brain mechanisms, and educators seeking to design educational materials and experiences to foster transfer need to build on a theory that is.

In the fourth paper (Chapter 5), I apply the analytical framework from the first paper to re-evaluate the position on neuroscience and education Bruer (1997) defines in "Education and the Brain: A Bridge Too Far." In addition, I define a different position on the issue, based on the observation that there is more than one possible bridge from neuroscience to education (e.g., the bridge described in this dissertation is distinct from the one Bruer describes), not all of which are equally far.

In the concluding chapter (Chapter 6), I place my research framework in the larger context of the emerging domain of educational neuroscience. Specifically, I argue that educational neuroscience does not have the structure of a formal discipline (like physics or mechanical engineering, for example), and therefore some other kind of theoretical infrastrcture is needed to faciltate work in the domain. I propose *design patterns* as a promising tool for addressing this problem. Design patterns are formal abstractions of successful solutions to recurring problems that have proven very useful in non-disciplinary domains such as architecture and software engineering. I also describe (and illustrate with several examples) how design patterns can be applied in educational neuroscience specifically.

To illustrate how each step of the research process can be executed and each tool applied, I work through a concrete example, starting with identification of a neural mechanism (CNDR) and carrying the analysis all the way through extraction of candidate educational principles. My intent is that this strategy of developing the general framework in parallel with a concrete applied case study serves multiple purposes: 1) it illustrates how each step of the research process connects with those that precede and follow it, 2) it demonstrates how the theoretical and analytical tools are applied to concrete questions and problems, 3) the entire case study represents a proof-of-concept that the neural-educational bridge can be crossed immediately, based on our current state of scientific knowledge, and 4) it provides a concrete case study of the complete research framework that potentially could be used as a pedagogical tool (for instance, to illustrate specific issues unique to this domain).

Chapter 2

Theoretical Framework: On the Relationship between Psychological Theories and Computational Models

Introduction

Fifty years ago, psychologists and computer scientists collaboratively invented a powerful framework for studying human psychology and behavior called the *symbolic paradigm* (Fodor, 1975, 1987, 1990; Gardner, 1985; Newell, 1980; Newell & Simon, 1976; Pylyshyn, 1986; Smolensky, 1988). This paradigm is based on the premise that the mind is fundamentally a symbol processing system, and that psychological research should therefore focus on characterizing its constitutive processes and how the symbolic data flowing through the system are transformed by them (for example, while solving problems). Researchers within this camp introduced the innovation of specifying psychological theories in terms of detailed data structures and algorithms that could (optionally) be loaded into a computer to simulate the psychological and behavioral processes entailed by the theory-program. For example, a cognitive model of problem solving in the game of chess could be specified as a computer program capable of playing chess using strategies gleaned from analyzing the behavior of human chess masters (Alden & Bramer, 1988; Bramer, 1982; Charness, 1992; Chase & Simon, 1988; Feigenbaum & Feldman, 1995). An early demonstration of the power of the symbolic paradigm was its central role in dislodging behaviorism as the dominant theoretical approach to studying human psychology and behavior (Chomsky, 1959; Gardner, 1985).

At around the same time, a competing computational framework based on a simple neural model called the "perceptron" emerged (Anderson & Rosenfeld, 1998; Gardner, 1985; Rosenblatt, 1958). Symbolic paradigm proponents eventually disposed of the perceptron as a viable basis for psychology using analytic methods (such as proof-by-counterexample) similar to those employed so effectively against behaviorism (Anderson & Rosenfeld, 1998; Gardner, 1985; Minsky & Papert, 1969). These dramatic dual demonstrations—against a purely behavioral paradigm in the first case and an ostensibly neural paradigm in the second—helped establish the symbolic paradigm specifically and cognitive science more generally as the dominant frameworks for studying psychology and behavior through the present day. This work also helped establish the "mind" as the predominant unit of analysis in psychological study.

About twenty years ago, researchers introduced a descendent of the original perceptron called the "multi-layer perceptron" (MLP)—also sometimes called the "connectionist model" (Anderson & Rosenfeld, 1998; Gardner, 1985; McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986). The MLP is superficially quite similar to the perceptron. The crucial difference is that the MLP overcomes the fundamental limitations that made the perceptron inadequate as a model of human cognition (Rumelhart, Hinton, & Williams, 1986; Widrow & Lehr, 1990). Smolensky (1988) has dubbed the psychological research framework based on these ANNs the *subsymbolic paradigm* to distinguish it from the symbolic paradigm. A key distinction between the two paradigms involves the theoretical role each accords to symbols. In the symbolic paradigm, the symbol is the fundamental unit of analysis. In the subsymbolic paradigm, in contrast, symbolic descriptions of cognitive structures and processes are

merely approximate discrete descriptions of a continuous underlying conceptual space (Smolensky, 1988).

In the last two decades the MLP and other kinds of artificial neural networks (or ANNs) have grown to challenge the symbolic paradigm as a framework for describing and investigating human cognition, although the precise nature and severity of that challenge have long been matters of debate (Elman, 1998; Smolensky, 1988). At one pole of the debate, some ANN proponents argue that these models represent a Kuhnian paradigm shift that could ultimately displace or absorb the symbolic paradigm (Churchland, 1981; Churchland, 1988; Granott, 1998; Lust, 2000; Schneider, 1987). At the other end of the spectrum, some members of the symbolic camp insist that ANNs simply offer an alternate implementation mechanism for their more abstract models, which has no direct bearing on the abstract symbolic models themselves (Broadbent, 1985; Fodor & Pylyshyn, 1988)—for example, in the way that statistical mechanics in physics is true at the "implementation" level but largely irrelevant when we are dealing with the Newtonian systems of structural engineering and everyday life (Smolensky, 1988). Based on the observation that many ANN strengths (e.g., perception, categorization) are weaknesses of symbolic models and vice-versa (e.g., symbolic models perform better on higher-order symbolic tasks like language and math), a third group has proposed "hybrid" models that integrate the two kinds of systems in an effort to leverage the capabilities of both in a single system (Just & Varma, 2002; Klahr & MacWhinney, 1998; Ohlsson, 2000; Sun, 1996; Tepper, Powell, & Palmer-Brown, 2002; Triantaphyllou, 2000; Wermter & Panchev, 2002).

In my view, progress in resolving this debate is hampered by the lack of an appropriate analytical framework for comparing and contrasting different psychological theories and models that respects the unique internal perspective of each paradigm while at the same time rendering them comparable via correlation with a shared external field of reference. For instance, the symbolic and subsymbolic paradigms are not directly comparable as they stand, because they are based on different assumptions, involve different units of analysis, employ different theoretical constructs, and are grounded in different frames of reference. Some kind of uniform code is needed into which they can both be translated for direct comparison to understand their similarities, differences, and unique characteristics.

I have three aims in this paper. First, I propose a meta-theoretical analytical framework that can be used to compare and contrast different psychological theories from the uniform perspective of philosophical materialism. Second, I apply this framework to map out the relationships between behaviorism, the perceptron, the symbolic paradigm, and the muli-layer perceptron to demonstrate the application and utility of my proposed framework. Third, I build on this inter-theoretical analysis to argue that the symbolic paradigm (grounded in a functionalist framework) entails at least two mutually exclusive materialist hypotheses about the nature of human cognition; these hypotheses cannot in principle be distinguished within the symbolic paradigm because of the strongly functionalist orientation of that paradigm; and the symbolic paradigm therefore lacks sufficient falsifiability to be considered scientific. I further argue that artificial neural networks (of which MLPs are just one example) represent an important advance over the symbolic paradigm because they commit to just one of these hypotheses, thereby

potentially exposing them to falsifiability in ways that make them more scientific than the symbolic paradigm as it is usually construed[1].

Analytic Framework

One difficulty arising in any attempt to compare and contrast different psychological theories is that each one seeks to parse the world in a different way. Indeed, the chief innovation of a new theoretical framework is often a novel unit of analysis and/or a different way of "cutting nature at its joints" to identify the boundaries of the entities under study. Piaget, for example, focused on the "epistemic subject" in his developmental theory (Gruber & Voneche, 1995), while Vygotsky (1986) took "word meaning" as his unit of analysis. Gardner (1993) identifies the external "domain" and the internal "intelligence" as meaningful units of analysis in his theory of multiple intelligences, whereas Fischer and Bidell (1998) make a different cut through some of the same phenomena in specifying the "skill" (which includes elements of both a person's internal ability and the external domain-specific task) and the "person-in-context" as theoretical primitives in their skill theory framework.

To complicate matters further, theories can differ along a large number of dimensions, including *a priori* assumptions, philosophical orientation (for example, materialist vs. idealist vs. dualist), methods employed, source and type of data used for empirical verification (and indeed whether empirical verification is even a consideration), level of analysis, granularity of description, phenomena of interest, etc. Even if we had adequate methods for adjudicating experimentally between the many theoretical proposals (which we do not, by and large), we would still first need to impose a *lingua*

[1] In Chapter 3 I describe my efforts to operationalize this potential for falsifiability in the form of an empirical experiment.

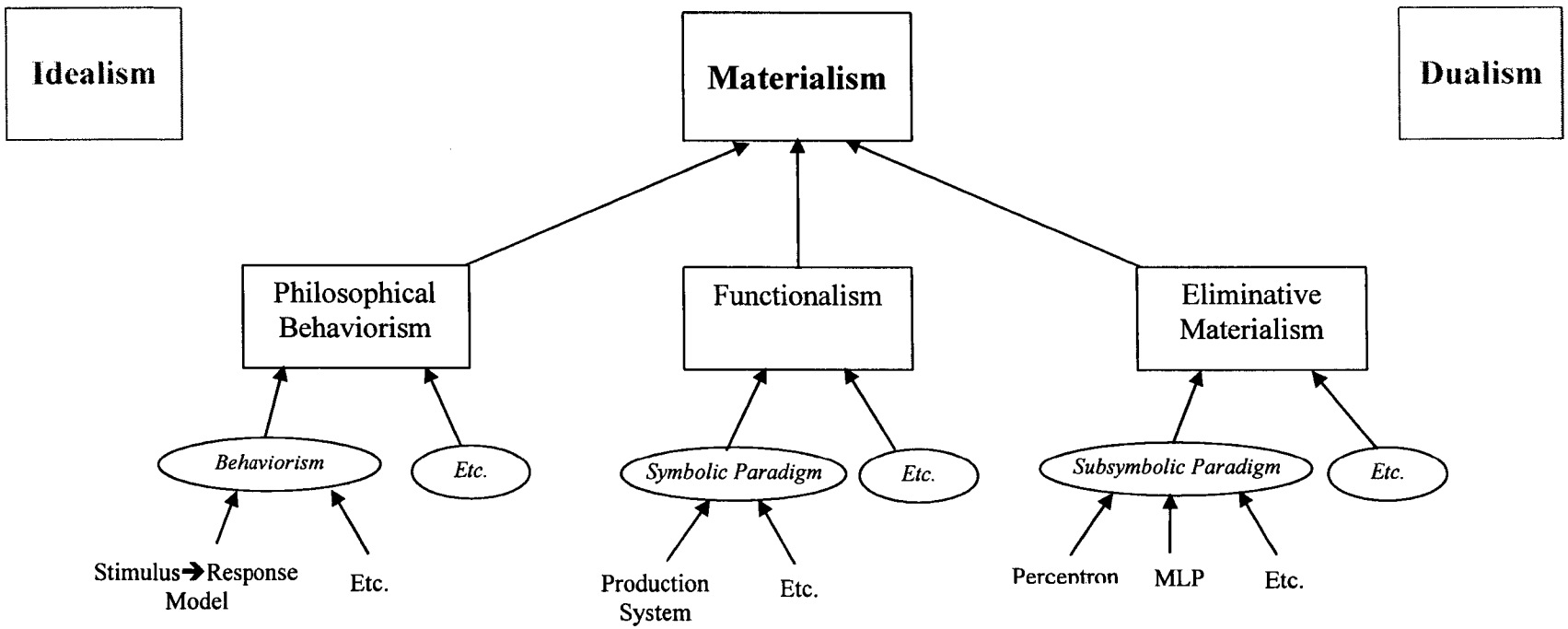
franca on this Tower of Babel tableau so we can at the very least determine when two parties are referencing the same entity in the world and when they are not.

Philosophical Materialism

In seeking to establish a common frame of reference for inter-theoretic analysis, I have found it useful to draw on the taxonomic system described by philosophers of science to organize different types of psychological theories (see Figure 2.1 for my take on this organization; also, see Churchland (1988) for a more thorough and very accessible treatment of these issues). At the leaves of the tree, there are many different specific *models* of psychology and behavior, including stimulus-response models, production system models, perceptron models, and MLP models. These can be grouped together under theoretical *paradigms* such as behaviorism, the symbolic paradigm, and the subsymbolic paradigm, as shown in Figure 2.1. A paradigm (loosely speaking) is defined by a set of assumptions, principles, and/or methods common to all the models that derive from it (Kuhn, 1996). Note that a paradigm can include multiple distinct kinds of specific models, as in the case of perceptrons and MLPs, both of which derive from the subsymbolic paradigm.

In the case of psychological theories, these paradigms can be grouped further into philosophical *camps* that cohere around a set of beliefs concerning how the brain-mind relationship should be handled theoretically. *Philosophical behaviorism* is based on the belief that mind and brain can and should be ignored for the purposes of developing a theory of human behavior (Churchland, 1988; Graham, 2002; Ryle, 1949; Watson, 1913). *Functionalism* entails the assumption that the mind can be studied independently of the brain (that is, that the brain can be ignored in developing a theory of human cognition—

Figure 2.1: Partial taxonomy of psychological models and paradigms grounded in philosophical materialism. This taxonomy explicitly excludes, for example, psychological paradigms, theories, and models derived from philosophical idealism or dualism.



Churchland, 1988; Levin, 2004; Putnam, 1975). *Eliminative materialism* holds that once a brain-level (or implementation-level) theory of cognitive processes is established, extant mind-level theories can be (and indeed will likely need to be) abandoned because they paint a "radically misleading" picture of cognitive functioning[2] (Churchland, 1981; Churchland, 1988; Churchland, 1986; Ramsey, 2003).

Finally, what all of these philosophical camps have in common is a shared basis in philosophical materialism. That is, most practicing scientists in these camps subscribe to the belief that behavioral and mental phenomena somehow ultimately derive from physical, chemical, and biological structures and processes located entirely in the body, and primarily in the nervous system (Churchland, 1988; Gardner, 1985). If these structures and processes were all accounted for, in other words, there would be no "residue" left over. The implication is that we do not need to include any additional factors—for example, ESP, magic, the soul, or a nonmaterial substance of mind such as that proposed by Descartes (1641/1960)—to explain cognitive phenomena. Materialism is differentiated in this way from dualism or idealism, for example, which allow for nonmaterial entities to influence the brain-mind system (and, by extension, behavior). The fact that all the theories under consideration in this paper are ultimately grounded in materialism suggests this as a promising point of departure for establishing a shared frame of reference, as I discuss in the next section.

---

[2] Although psychological behaviorism would seem to imply a commitment to philosophical behaviorism and the mainstream symbolist philosophy is functionalist, I do not mean to suggest that all researchers working with ANNs are necessarily eliminative materialists. Figure 2.1 is meant to illustrate four points: 1) there is a great diversity of extant psychological models, theories, and paradigms; 2) these frameworks tend to cluster into larger groups as you move up the tree; 3) the common ground shared by many of these diverse frameworks is a belief that the brain somehow gives rise to the mind (philosophical materialism); and 4) in the present analysis I am interested only in models, theories, paradigms, and camps that are grounded in this materialist view (this excludes, for example, theories grounded in idealist and dualist philosophies).

Criteria for the Shared Analytic Framework

Materialism provides a starting point for my general analytic framework for understanding relationships between psychological theories in two ways. First, materialism delineates unambiguously the boundaries of the system under study, and establishes the upper limit on what needs to be included in it. This includes the nervous system (and to a lesser extent the body housing it), with all of its physical, chemical, and biological structures and processes. The mind and behavior are in this view recognized as being in a meaningful sense redundant (although possibly indispensable) descriptions of the nervous system, arising as they do from these underlying entities. These two levels of analysis must nonetheless be included independently in the analytic framework to accommodate specific theories like behaviorism (which seeks to construct a theory on behavioral data alone, without reference to the mind or brain) and the symbolic paradigm (which assumes the mind can be described without specific reference to the details of the nervous system).

Second, materialism suggests a useful way to parse the system—in terms of the physical and measurable structures and processes at each level of analysis. That is, the behavioral level of analysis will be defined strictly in terms of measurable behavioral phenomena (stimuli and responses), the brain level of analysis will be defined strictly in terms of measurable physical brain structures like synapses, etc.

To these guiding principles derived from materialism, I add constraints based on the purposes for which I want to use the framework. I want to use this framework as an unambiguous frame of reference for understanding how different theories relate to one another. It seems prudent, therefore, to define the levels of analysis so that they facilitate "bookkeeping" of the theoretical constructs entailed by a psychological model or theory.

For instance, a single theoretical construct (e.g., working memory) should not legitimately belong at two levels of analysis (for example, mind and brain), because then it would be "counted" twice[3]. This strategy will support unambiguous inferences and conclusions about the theories under consideration.

To summarize, in constructing my analytic framework, I sought to meet four criteria:

1) **Materialist Criterion**: Elements at each level of analysis should be meaningfully identifiable with physical/measurable phenomena. For example, at the behavioral level, inputs can be identified with measurable stimuli (such as visual images, auditory prompts, or tactile inputs) and outputs can be identified with measurable responses (such as reaction time, categorical response, or eye blinks).
2) **Uniqueness Criterion**: The levels of analysis should be specified to minimize overlap between them. For example, if a raw sensory stimulus is accounted for at the behavioral level of analysis, then it should not also be included at the "mind" or "brain" level of analysis unless: a) there is some physical referent at the other level to which it can be attached, such as a neural representation (from criterion #1), and b) there is some account of the relationship between the two physical referents at different levels, such as a specification of the transformation from raw input to neural representation. This facilitates careful "bookkeeping" by making sure each theoretical construct is accounted for *at most once.*
3) **Completeness Criterion**: Levels of analysis should accommodate all significant phenomena relating to brain, mind, and behavior. This criterion facilitates

[3] Although different parts of the working memory subsystem might be placed at different levels of analysis, the point is that no single entity should appear in more than one place.

bookkeeping by making sure that each relevant theoretical construct can be accounted for *at least once.*

My focus on theoretical paradigms with a shared foundation in philosophical materialism helps to establish an upper bound on the phenomena that need to be accommodated in this framework to satisfy criterion #3. In addition, note that criteria #2 and #3 taken together help to ensure that any theoretical construct will be accounted for *exactly once* in the analytic framework—to the best of my knowledge a unique feature of the current proposal. These first three criteria constitute a set of definitive requirements on the analytic framework. The final criterion is more heuristic, intended to guide general decision making (such as how many levels of analysis to include) and to facilitate the application of this analytic framework to diverse psychological models and paradigms:

4) **Translation Criterion**: When possible, use levels of analysis bearing some relation to other levels-of-analysis frameworks in wide use. For example, people often talk colloquially in terms of "brain," "mind," and "behavior"; cognitive scientists often rely on Marr's "computation," "representation," and "implementation" levels of explanation (Marr, 1982; Posner, 1989); and others commonly use domains or disciplines like "education," "cognitive psychology," and "neuroscience" dealing with phenomena at different levels of organization as stand-ins for formal levels of analysis (see, for example, Bruer, 1997).

Defining the Levels of Analysis

In accordance with criterion #4, as a starting point for defining the three levels of analysis, consider the colloquial terms "brain," "mind," and "behavior." Behavior can be

defined simply as any directly observable and/or measurable externalized action (including such experimentally elicited responses as linguistic utterances, button presses, and eye movements), or any stimulus applied to the body that can be registered by the senses (such as pressure, temperature, pain, visual images, or sounds).

The word "brain" is most closely associated with the pinkish organ situated inside the skull—the complex structure composed of smaller structures like cells, synapses, and proteins.

As a first step, and in keeping with constraint #2 (non-overlapping levels of analysis), the mind can be defined in terms of the other two—roughly speaking, it is everything that comes "between" the physical organ of the brain and the externally observable behavior. That is, "mind" is an abstract category containing all the internal representations and processes not directly observable that enable behavior and that are ultimately instantiated physically in the brain.

These definitions of brain and mind raise a difficulty, however. The brain is most closely associated with the physical organ by that name, but the brain also has a functional aspect. Structures like cells, synapses, and neurotransmitters support the generation of activity patterns like neural action potentials (or "spikes"). In particular, when a neuron becomes sufficiently stimulated, it generates a series of action potentials called a "spike train" that encodes information. For example, a single neuron in the cat visual system might produce many spikes per second when the animal views a horizontal line but almost no spikes in response to a vertical line (Hubel, 1995). Collectively, many neurons sensitive to different stimulus features (color, line orientation, movement,

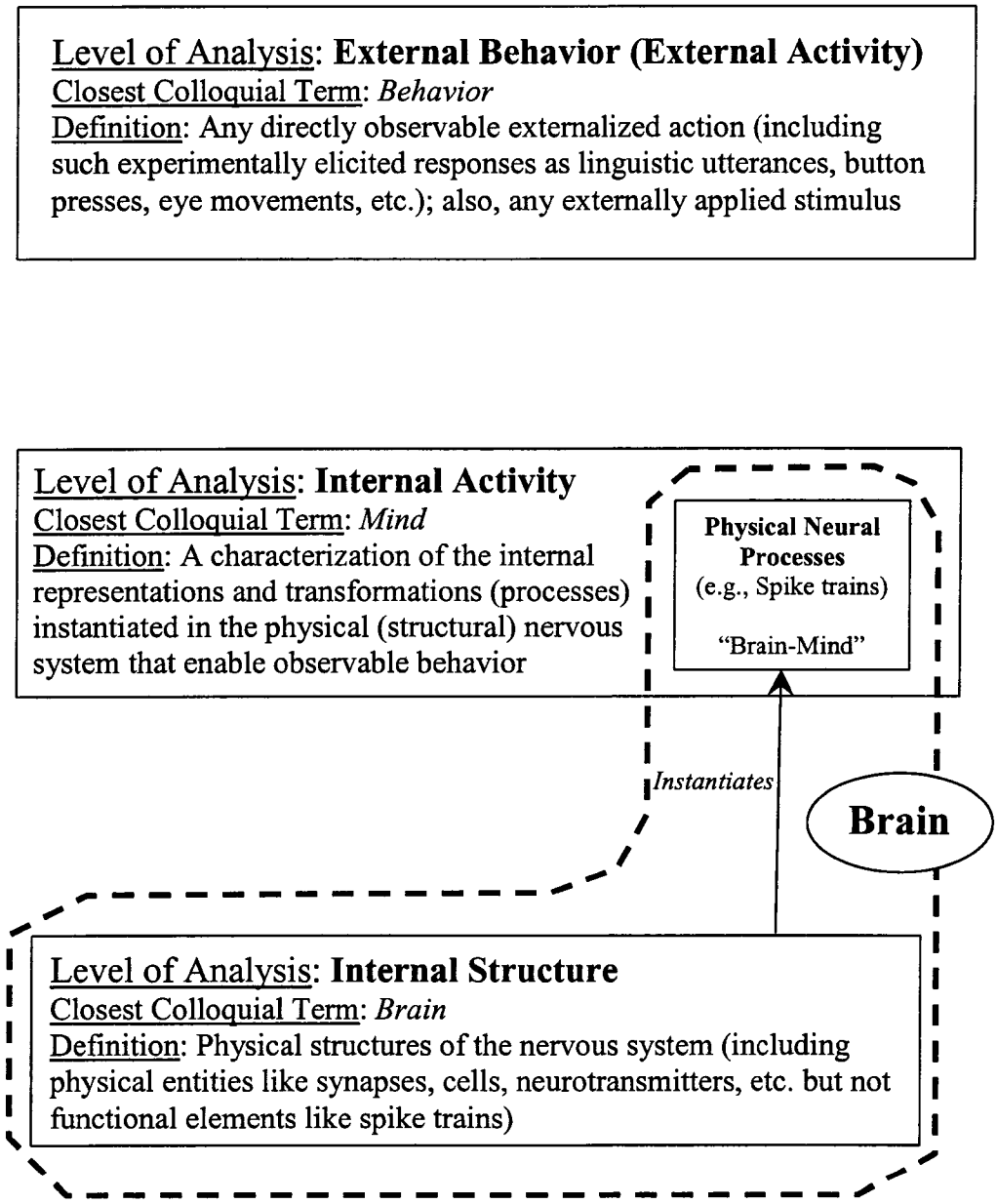
brightness, etc.) encode information about the structure of the visual scene in the spike trains they produce.

In order to understand the difference between structures (such as synapses) and activity patterns (such as spike trains), imagine scientists could flash-freeze a fully functional human brain without damaging it, simultaneously cutting off its energy supply and blocking its sensory inputs so all activity would cease completely. All the components of the brain that can be observed while the brain is in this frozen, inactive state (including synapses, cells, and neurotransmitters) are structures. All the phenomena that existed while the brain was active but disappeared at the moment it was frozen (including action potentials and spike trains) are activity patterns.

Activity patterns like spike trains are measurable physical phenomena, and in that sense they should be considered part of the brain. However, these phenomena are information-carrying processes (or the products of processes), not independently stable material structures like cells, and therefore they also participate in the mind category. We can refer to these physical brain processes collectively as the "brain-mind" to distinguish this description of the mind from alternative descriptions derived from other sources of data, such as behavioral observation.

For my purposes, the distinction between physical entities like synapses and spike trains on the one hand and functional categories like "mind" and "behavior" on the other hand is as important as the distinction between the levels of analysis, so I introduce my own nomenclature to preserve both (see Figure 2.2). Instead of the functionally defined, ambiguous, and overlapping categories "brain" and "mind," I introduce the materially grounded (constraint #1) and non-overlapping (constraint #2) categories "internal

Figure 2.2: Levels of analysis defined



structure" and "internal activity." According to the materialist doctrine, any behavioral or cognitive structure or process must ultimately have its physical basis in either nervous activity or nervous structure, so this framework should in principle be able to accommodate virtually any theoretical construct grounded in a materialist paradigm. This scheme therefore satisfies all four of my design constraints.

Previous Levels of Analysis Proposed for Analyzing Brain-Mind-Behavior Relationships

Before jumping into the analysis itself, let me mention briefly why I felt it necessary to introduce a whole new framework rather than using one that has already been proposed formally or used informally in the past. The short answer is that no other analytic scheme with which I am familiar meets all four criteria laid out above, the need for each of which I have tried to justify explicitly. I will demonstrate with three familiar examples.

Most commonly, these issues are discussed informally in terms of "brain," "mind," and "behavior"—this is certainly the usual way of discussing them in the popular press, for example. I described above how the "brain" and "mind" categories are overlapping (Figure 2.2 illustrates this), which violates design constraint #2 (uniqueness criterion).

Another common approach is to use disciplines to stand in for the three levels of analysis. For example, Bruer (1997) conducts an analysis of neuroscience and education using the following rough associations: neuroscience ~ brain, cognitive psychology ~ mind, education ~ behavior. The problem here is similar to that in the previous case. Specifically, "neuroscience" encompasses many different types of data, methods of analysis, and levels of organization, some of which encroach on the behavioral and/or

mental planes. Cognitive neuroscience, for example, focuses on functional patterns of neural activity. Presumably, this neural activity overlaps with the physical referents behind cognitive psychological models of cognitive processes associated with specific behaviors. Therefore, cognitive neuroscience and cognitive psychology are not suitable as levels of analysis because they violate the uniqueness design constraint. Because many disciplines span multiple levels of analysis in this way, they would also violate design constraint #2[4]. This scheme also violates constraint #3 (completeness criterion) because it is often difficult to reconcile data and methods from disciplines (like computational neuroscience) not included in a study with those that define the levels of analysis used there, which means there is no guarantee that all relevant phenomena can be accommodated by any specific set of disciplines. These inherent differences—which are, in many cases, incommensurabilities—are largely what define disciplinary boundaries in the first place and thus make disciplines a poor basis for defining levels of analysis.

The formal framework used most widely in the cognitive science literature is that proposed by Marr (1982). His three "levels of explanation" are called *computational theory*, *representation and algorithm*, and *hardware implementation*. I apply my framework to examine Marr's framework in detail in a later section and to argue that it violates at least one of my criteria, which is why I felt the need to introduce a new framework rather than using his.

In summary, all of these popular schemes can be rejected based on the fact that they employ overlapping categories, which means that sometimes a single entity can be identified with more than one level, which makes it difficult to reason carefully about

[4] In chapter 5 I examine this particular example in greater depth.

theories (especially in relation to one another). In addition, ambiguous specification of the different levels of analysis in each scheme frequently makes it difficult to know where to place a phenomenon under study (for example, any example of brain activity itself—should it be considered part of the brain or mind, or perhaps both?). My framework subsumes key features of all of these other schemes and replaces them with one uniform analytic framework grounded in physical structures and processes.

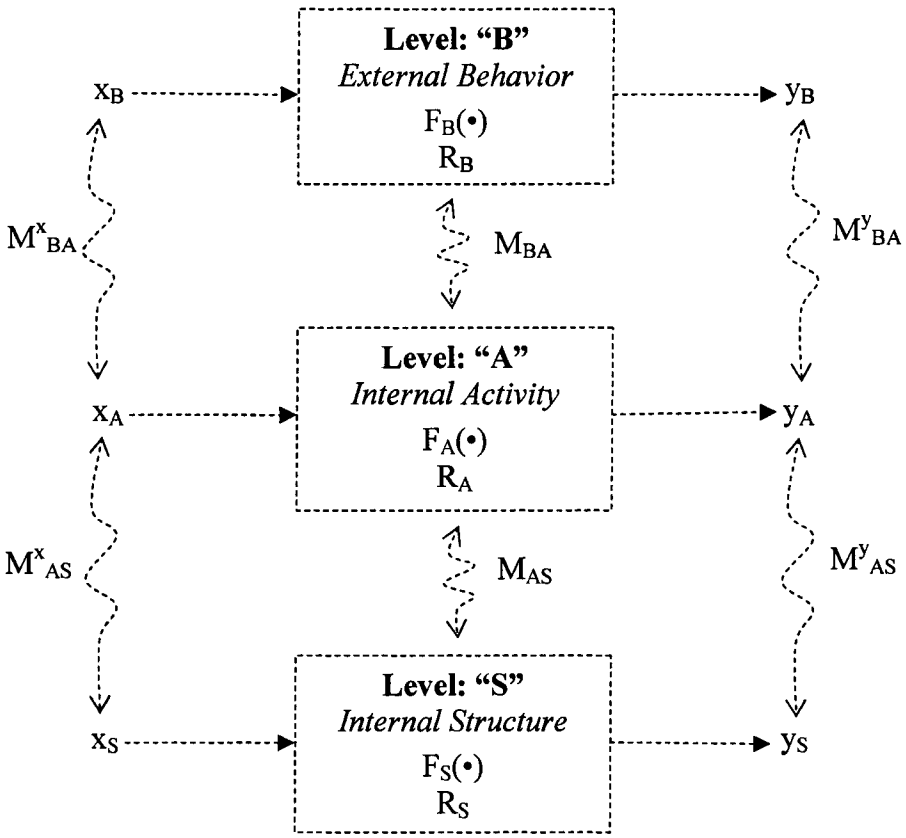
The Complete Analytic Framework

The levels of analysis form the core of the analytic framework, but by themselves they are not sufficient for present purposes. In addition, we need some way to track the measurable information as it passes through and is transformed by the system, and to relate specific structures and processes appearing at different levels of analysis to each other. Therefore, I have added a number of elements to the core framework, as shown in Figure 2.3. These particular elements are added with the goal of preserving important *within-level distinctions* (for example, by differentiating external inputs and outputs from internal representations at each level) while also explicating *cross-level mappings* (for example, by specifying how a description of an externally applied stimulus relates to a description of the same stimulus in terms of a pattern of internal activity). These additional features facilitate careful bookkeeping of theoretical elements.

Wihin-Level Distinctions

Each level of analysis *i* has its own set of inputs (x_i) and outputs (y_i). Each level of analysis also specifies a set of transformations from its inputs to its outputs, and the functions specifying this set of transformations for level *i* are specified collectively as F_i. In addition, each level can potentially store information as representations (R_i).

Figure 2.3: The proposed analytic framework is based on three levels of analysis. Each level (*i*) is specified as a set of inputs (x_i), outputs (y_i), representations (R_i), and functions (F_i). Relationships between levels *i* and *j* are specified as mappings between the levels' inputs (M^x_{ij}), the levels' outputs (M^y_{ij}), and the levels' internal mechanisms (M_{ij}).



For example, an electric shock applied to a subject's hand would be a behavior-level input (x_B). If the subject flinches in response to the shock, the flinch would be identified as a behavior-level output (y_B). A functional description of the relationship between the shock stimulus and the flinch response that makes no reference to internal representations would be a transformation identified as an element in the set of behavior-level input-output transformations (F_B). If the subject later makes an entry in his journal about the experience of being shocked, this written record would be identified as a behavior-level (or external) representation (R_B).

Cross-Level Mappings

Relationships between inputs at different levels are denoted by M^x_{ij}, which stands for "mapping between the inputs at levels *i* and *j*." Similarly, relationships between outputs are denoted as M^y_{ij}. Relationships between the internal mechanisms at levels *i* and *j* (especially between their internal representations and transformation functions) are denoted by M_{ij}.

For example, imagine an experiment in which a monkey is shown a printed word (e.g., "cup," "ball," or "banana") and is expected to point to the indicated object (a cup, ball, or banana, respectively) included in a set of objects before it. Imagine further that an array of microelectrodes has been implanted surgically in the visual centers of the monkey's brain to record the spiking activity of one thousand individual neurons for the duration of the experiment. The behavior-level input (x_B) in this scenario is a word (e.g., "cup"). When presented to the monkey, each word induces a unique pattern of neural activity as measured by the microelectrodes, and this pattern of neural activity corresponds to the input at the internal activity level (x_A) in this scenario. A table

specifying the correspondence between the printed words and their associated neural activity patterns would constitute the mapping (M) between inputs (x) at the behavioral (B) and activity (A) levels (M^{x}_{BA}).

Integrated Framework and Analytical Strategy

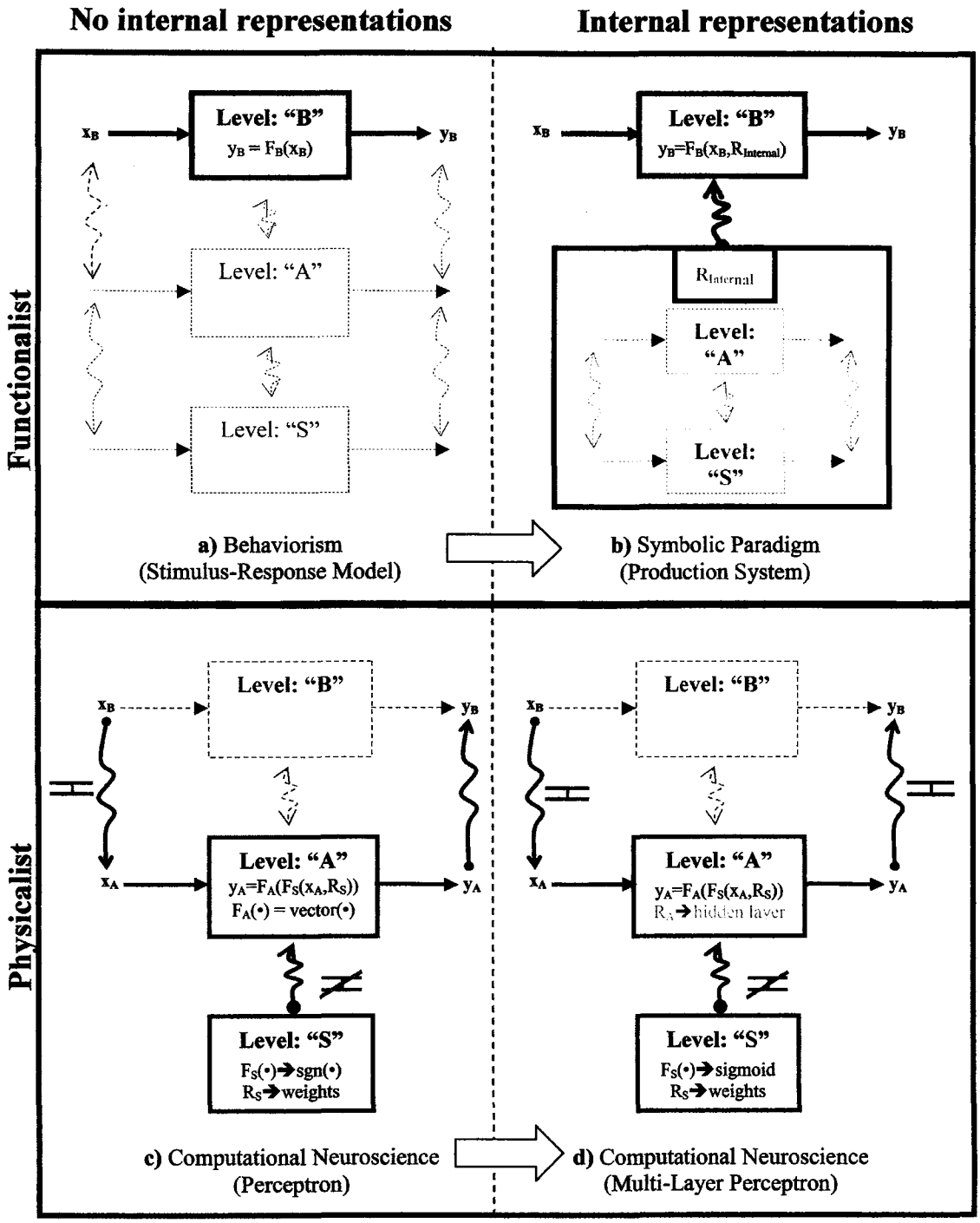
To summarize, each level of analysis *i* is defined by a quadruple of elements (x_i, y_i, R_i, F_i), and relationships between any two adjacent levels *i* and *j* are specified by three mappings (M_{ij}, M^{x}_{ij}, M^{y}_{ij}). By contrast, other frameworks in this domain (such as those described above) tend to define the levels of analysis ambiguously and monolithically (that is, without differentiating the kinds of elements operating at each level). The goal here is to include enough detail in the framework to highlight significant differences between different models or theoretical paradigms while also suppressing *unnecessary* detail so as to illuminate meaningful similarities between them.

My approach involves mapping disparate theoretical frameworks onto this materialist framework so they can be compared in a uniform manner. In the following sections I describe in general terms how the framework can be applied. I then demonstrate this process concretely by applying it to behaviorism, the symbolic paradigm (using the production system as a concrete exemplar), perceptrons, and MLPs. The product of these four analyses will be the taxonomy previewed in Figure 2.4, which I use to compare and contrast the four different theoretical frameworks.

Intertheoretic Analysis

The application of the analytic framework to a specific theory or model involves four steps (although not necessarily applied in the order presented). First, identify any

Figure 2.4: Taxonomic summary of the analyses of four psychological models: a) behaviorist stimulus-response model, b) symbolic paradigm production system model, c) single-layer perceptron, and d) multi-layer perceptron (MLP).



measurable data modeled or otherwise incorporated into the theory, and associate them with the (x_i) and outputs (y_i) at the appropriate level of analysis. For example, external behavior would obviously appear as inputs and outputs at the top level ("B"), while single-cell neural recordings measure internal activity, so they would be identified with inputs and outputs at the middle level ("A")[5]. Second, identify the functions (F_i) that transform inputs into outputs, and (usually) locate them at the same level of analysis as the inputs and outputs that they relate (but see the perceptron and MLP analyses for two examples where a basis function is at a different level from its inputs and outputs). For example, a behaviorist model relates external stimuli to external responses without reference to anything internal, so it would be identified as a set of functions at the behavioral level (F_B). Third, identify the representations stored at each level of analysis. Fourth, identify the mappings that relate previously identified elements located at different levels to one another. I demonstrate this procedure concretely in the following sections.

Behaviorism

Behaviorism is a psychological paradigm emphasizing externally observable (that is, behavioral) aspects of thought. Although there have been many strands of behaviorism (Graham, 2002; Hatfield, 2002), including some that have appealed to mentalistic or quasi-mentalistic entities such as internal "mediating" variables or processes (Berlyne, 1965; Graham, 2002; Hatfield, 2002; Osgood, 1953), a number of early behaviorists (most notably Skinner, with his self-styled "radical behaviorism") sought to explain behavior entirely in terms of measurable behavioral responses to

[5] In cases where a model is theoretically rather than empirically based, this information can be inferred from the description of the elements the theory or model is intended to represent.

measurable stimuli, without reference to internal mental states, neurological activity, or neurological structure (Graham, 2002; Skinner, 1938a, 1945; Watson, 1913). Many of these early behaviorist models were based on the stimulus-response framework of classical conditioning (Pavlov, 1927). Although it is straightforward in this case to locate the behaviorist stimulus-response model within my analytic framework directly (see Figure 2.5), for demonstrative purposes I now apply the four-step process described above to show how a theory can be mapped systematically onto the framework.

Step #1: Identify the data source upon which the model is based

The basic structure of a stimulus-response model is:

$$\text{Stimulus} \rightarrow \text{Response}$$

Alternatively, this can be written another way:

$$\text{Response} = F(\text{Stimulus})$$

It is particularly easy to identify the source of relevant data in the case of behaviorism, since by definition these theorists sought to base their theories entirely on measurable behavioral stimuli (x_B) and measurable behavioral responses (y_B). Rewriting the equation to reflect this yields:

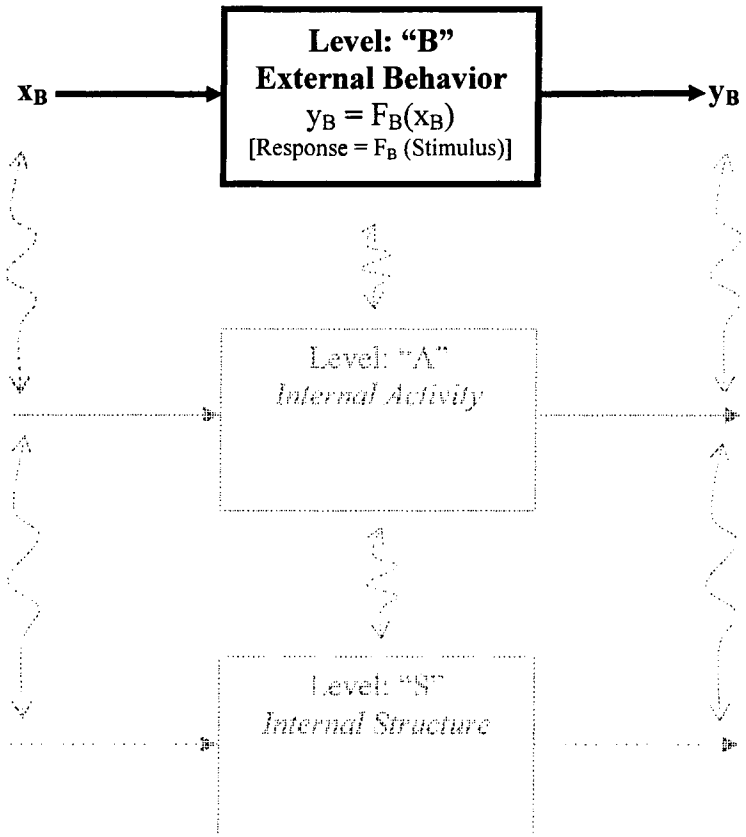
$$y_B = F_i(x_B)$$

Step #2: Identify the functions relating inputs to outputs

Since the only inputs and outputs are at the behavioral level of analysis, and the behaviorists specifically sought to avoid reference to internal processes and structures, the functions relating these inputs and outputs therefore also have to be at the behavioral level (F_B):

$$y_B = F_B(x_B)$$

Figure 2.5: A stimulus-response model from behaviorism represented in terms of my analytic framework



Step #3: Identify any representations (other than inputs and outputs)

Because of the focus on behavior alone, we can remove everything below the first level of analysis from the diagram. The only possible representations would therefore be at the external/behavioral level (R_B). Behaviorism allows for no representations (as distinct from input-output functions) other than the inputs and outputs, so we remove R_B from the diagram.

Step #4: Identify the mappings between levels

There is only one level active in behaviorism, so there are no mappings between levels.

The main focus of research and theory in radical behaviorism was on determining the nature of $F_B(\bullet)$—that is, determining the contingencies between a set of measurable stimuli and a set of measurable behavioral responses. This is a useful model for a range of animal behaviors (Jennings, 1906; Skinner, 1938b), as well as some human behavior (Leitenberg, 1976; Rincover, Newsom, Lovaas, & Koegel, 1977; Stahl, Thomson, Leitenberg, & Hasazi, 1974), and had it panned out generally, it would have been a major triumph for scientific psychology. Unfortunately, it was insufficiently powerful to serve as the basis for a comprehensive theory of human psychology and behavior, as the cognitivists demonstrated (Chomsky, 1959; Gardner, 1985; Graham, 2002).

The fundamental problem with the stimulus-response model can be identified in step #3 above—it lacks any capability for storing internal state (internal representations). For example, if a person faced with a challenging problem tries one strategy unsuccessfully five times in a row, she might decide on the sixth time to switch strategies based on the observation that the first one is not working. There is no information in the stimulus that triggers this difference—the stimulus situation could be identical in all six

trials. Instead, the difference is caused by the fact that the person is *internally* keeping track of her past tries and using that information as part of her strategy-selection criterion. Since this internal information is not readily (or necessarily) accessible behaviorally (e.g., in the stimulus), the behaviorist stimulus-response model is unable even in principle to account for this important source of systematicity in the subject's behavior. I have described a trivial example to demonstrate the general point. Chomsky (1959), Lashley (1951), and others have described more involved examples relating to linguistic behavior and other domains, but these examples all point to the same root problem—no capacity for storing internal state (representations) in the stimulus-response model.

Cognitivism: Production System

In the 1950's the cognitivists emerged on the scene and demonstrated definitively that the radical behaviorist formulation would not work as a general theory of human cognition. The reason, as I already mentioned, is that there are important phenomena (such as aspects of language) that can be shown to require internal storage in order to be explained adequately. In essence, the cognitivists demonstrated unequivocally the need for some kind of internal representations—which are often identified with the "mind" that the behaviorists had tried to suppress—in any adequately complete theory of human psychology and/or behavior.

To see how the cognitivist framework differs from the behaviorist, consider the most prominent type of model within the symbolic paradigm—the production system (Anderson, 1987, 1993; Klahr & MacWhinney, 1998; Newell & Simon, 1961, 1972; Simon, 1992, 1999). A production system is a computational cognitive model that has two interacting components: working memory and production memory (Klahr &

MacWhinney, 1998; Simon, 1999). Working memory stores a set of elements representing the current state of the world and/or the results of previous internal processing. For example, a production system for solving a balance beam task would include features of the balance beam problem, such as number of weights on each side and their distances from the fulcrum, in its working memory. Production memory consists of a set of condition-action rules called productions. In the balance beam production system, these rules would determine how the weights and distances encoded in working memory are transformed into predictions about which side of the balance beam will go down. When the condition part of a production matches an element in working memory, the action part is executed, which in turn might modify working memory and/or produce an output from the system.

Measurable Data

The standard approach to developing a production system model is an experimental procedure known as the "talk-aloud protocol" (Newell & Simon, 1961, 1972). The basic idea is to pose a series of representative problems in the target domain to a group of human subjects (often, but not necessarily, domain experts) and have them externalize (through verbalization, writing, etc.) their thought processes as they solve the problems. Researchers analyze these data to identify patterns that inform the specification of algorithms (productions) and data structures (contents of working memory) that parsimoniously summarize the observed human performance data. The external inputs to the model (x_B) are representations of key elements of the problems, and the model outputs (y_B) are the final products of the problem solving sessions. The measurable data on which such models are typically based is therefore entirely behavioral.

Representations (other than inputs and outputs)

In addition to the model inputs and outputs, the model also produces intermediate products (e.g., partial solutions, temporary representations of sub-problems) that do not appear in the inputs but are derived from them, and that are used to produce but do not necessarily appear in the final solution that is identified as the output. These internal representations ($R_{internal}$) are stored alongside the inputs and outputs in working memory, but in the present analysis these internal representations are distinguished from the system inputs and outputs[6].

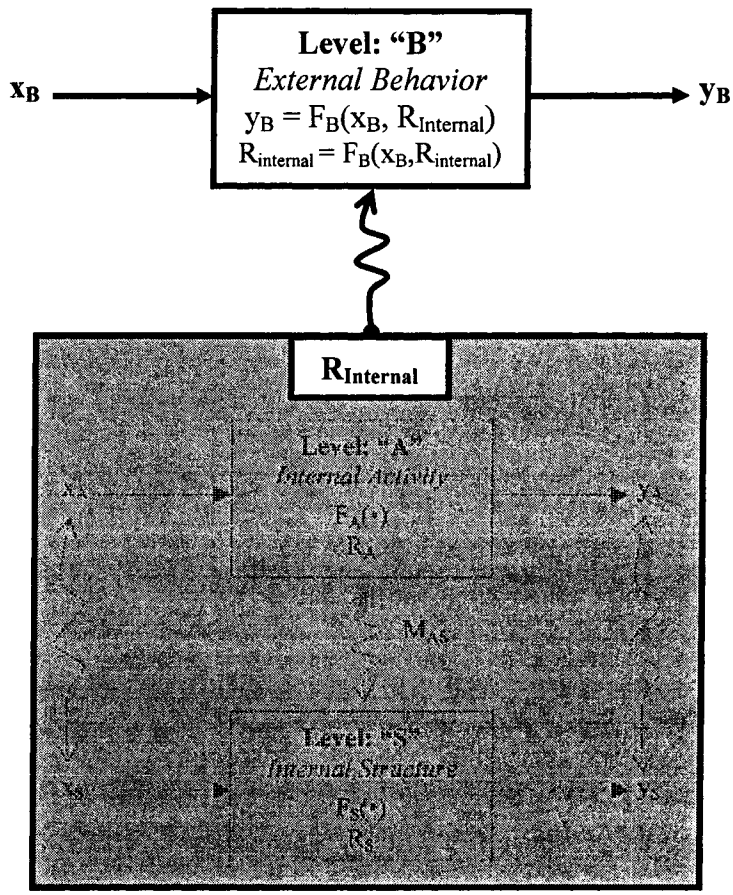
The question is, where should these representations be placed on the analytic framework? They are definitely *internal* representations, so we can rule out R_B. But the cognitivists are explicitly agnostic with regard to the physical structures and processes supporting these internal representations (Marr, 1982; Newell & Simon, 1972; Simon, 1992; Smolensky, 1988); in fact, this is an important defining feature of the functionalist view (Churchland, 1988; Levin, 2004; Smolensky, 1988). They are, however, materialists, so we can at least state with certainty that the internal representations in this theoretical framework can in principle be described by some unknown function of structural and activity-based functions and representations in the nervous system:

$$R_{Internal} = F_?(R_S, R_A, F_S, F_A)$$

This is depicted in Figure 2.6 by combining the activity and structural levels (and all their inputs, outputs, and contents) into a single black box labeled $R_{internal}$.

[6] Note that this demonstrates a unique feature of the current analytic framework: elements that look superficially the same within a modeling paradigm—such as the diverse elements stored in a production system's working memory—are identified systematically in a more differentiated manner here. In this case, the contents of working memory are identified as either inputs, outputs, or the intermediate products of problem solving processes. An important feature of this framework is that it supports a separation between the physical model and the theory it represents.

Figure 2.6: A production system model from the symbolic paradigm represented in terms of my analytic framework



Functions Relating Inputs to Outputs

A generic production has the form:

If **<Condition>** then **<Action>**

Or, rewriting it as a function:

$$\textbf{<Action>} = F_i(\textbf{<Condition>})$$

The production system inputs and outputs are derived from behavioral data, and the internal representations are inferred from behavioral data without regard for the underlying organization of physical structures and processes. Similarly, the productions are inferred from behavioral data with the goal of providing a parsimonious account of the transformation from behavioral inputs to behavioral outputs, without reference to specific physical structures and processes internal to the system. I would argue, therefore, that the functions relating inputs (x_B) to outputs (y_B) also belong at the behavioral (or external) level (F_B):

$$\textbf{<Action>} = F_B(\textbf{<Condition>})$$

Mappings Relating Elements Across Levels

On the surface, the production system so far looks very much like the behaviorist's stimulus-response model from the previous section:

$$\textbf{<Response>} = F_B(\textbf{<Stimulus>})$$

The difference, of course, is that the productions link external behavior to internal representations. Specifically, the production conditions apply to internal representations as well as system inputs, and the production actions can modify internal representations as well as producing outputs:

$$R_{Internal} = F_B(x_B, R_{Internal})$$
$$y_B = F_B(x_B, R_{Internal})$$

Note that these equations also implicitly define M_{BA}, which specifies the relationship between elements at the behavioral and internal activity levels.

If we step back and look at the system in terms of input-output behavior alone, we can ignore the first equation and characterize the system using just the second one:

$$y_B = F_B(x_B, R_{Internal})$$

This looks superficially like a small difference from the behaviorist stimulus-response model (involving just the addition of $R_{Internal}$), but the differences in terms of implications could hardly be more dramatic. The behaviorist framework is demonstrably too weak to model certain straightforward but important aspects of human behavior (Chomsky, 1959; Gardner, 1985; Jeffress, 1951; Lashley, 1951). The symbolic paradigm, in contrast, is in principle powerful enough to model any input-output behavior, no matter how complex (Chown, 2004). The potential power of these models is not a problem *per se* and, indeed, is often cited as a point in their favor. In a later section I will argue, however, that this degree of power combined with the functionalist insistence on agnosticism regarding implementation details insulates the symbolic paradigm from empirical verification. That is, these two factors combined basically render symbolic paradigm models virtually unfalsifiable, thereby moving them beyond the reach of scientific investigation.

Computational Neuroscience: Perceptrons

Around the same time as the cognitive revolution was gaining momentum, people in the nascent field of computational neuroscience were exploring neurally-inspired models of information processing in an effort to understand how cognitive functions are physically implemented in neural structures and processes (Anderson & Rosenfeld, 1998;

Sejnowski, Koch, & Churchland, 1988). One of the most famous early neural models was called the perceptron (see Figures 2.7 and 2.8; Anderson & Rosenfeld, 1998; Gardner, 1985; Rosenblatt, 1958). Researchers sought to use the perceptron model to gain insight into the basic neural mechanisms of information processing underlying more complex forms of perception and cognition.

Mathematical Definition of the Perceptron

The perceptron was intended to model (however crudely) the behavior of neurons (Figure 2.7) or collections of neurons in the nervous system (Figure 2.8; Anderson, 1995; Rosenblatt, 1958). For a network of units working together on a single problem, the j^{th} node computes a simple function of its inputs and associated weights (w_i) of the following form:

$$output_j = sgn(\Sigma_i\ w_i * input_i)$$

In this equation, sgn(•) is a function returning the sign of its argument (-1 or +1), and the index *i* ranges over all the inputs for the node under consideration.

If there is more than one node in the network, then the inputs to the network are just the inputs to all of its constituent nodes, and the network output is simply the vector of outputs for all the nodes in the network (Figure 2.8):

$$output_{Network} = vector_{all\_j}(output_j) = vector_{all\_j}(sgn(\Sigma_i\ w_i * input_i))$$

The perceptron can be analyzed by systematically identifying how each element in the equation above can be identified with an appropriate element in the analytic framework of Figure 2.3.

Figure 2.7: Key structures of a spinal motor neuron (left), and corresponding elements of an analogous node from a perceptron (right). In the biological neuron, dendrites collect stimulation from other neurons. If the total stimulation arriving on all the dendrites at one time exceeds a threshold, the neuron fires, sending activation down its axon. The perceptron node performs a similar operation. It sums its inputs, and if that sum is greater than zero it "fires" by generating a positive output. The efficacy of the biological synapse (which is modified by biological learning processes) is represented in the simulated model by the weight on each input connection (which is modified by simulated learning algorithms). The inputs to a neuron are modulated by the synaptic efficacy before being summed at the cell body, just as the input to a perceptron node is multiplied by the weight before being summed by the node.

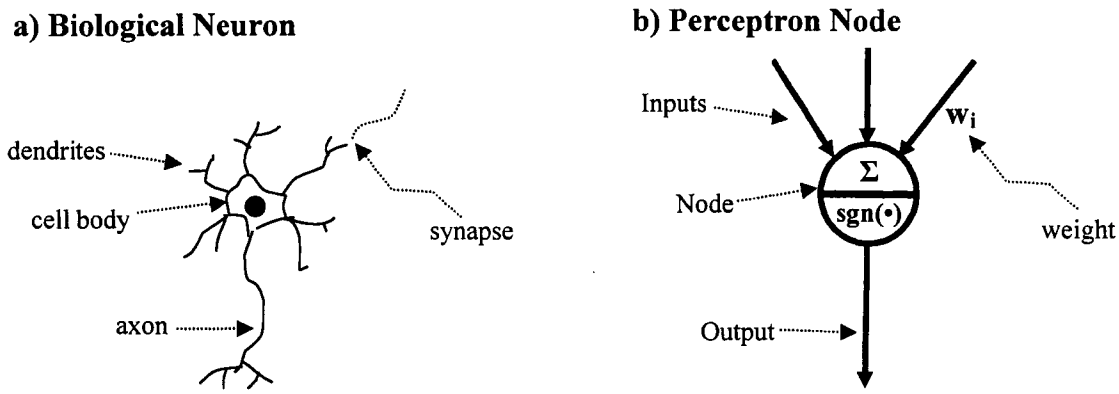
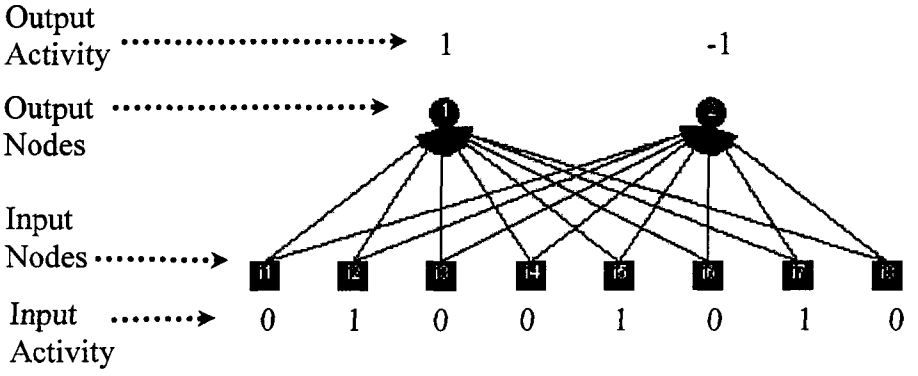


Figure 2.8: Example of a perceptron network. The original perceptron had one layer of modifiable connections between input and output nodes.



First, the inputs and outputs for each node are meant to model patterns of activity in the nervous system, so these can be identified with x_A and y_A (dropping the node subscripts for readability and rewriting the argument to sgn(•) as a list of variables):

$$y_A = \text{sgn}(w, x_A)$$

The perceptron weights are meant to model structures in the nervous system like neuron thresholds and synapse strengths, so these can be identified with representations at the structural level (R_S):

$$y_A = \text{sgn}(R_S, x_A)$$

The sgn(•) function is a model of the transfer function of a basic neural computing element (whether an individual neuron, a cortical column, or some other basic unit of computation in the nervous system). If we assume for the sake of concreteness a correspondence between a single neuron and a perceptron node, then I would argue that the node transfer function should be identified as a function at the structural level (F_S), since this is a basis function embodied in the neural structure that does not necessarily relate directly to the activity-level function computed by the system. In other words, at the network level the perceptron might be categorizing visual images, making medical diagnoses based on symptom patterns, or computing a whole host of other functions, but this basis function maintains its functional form regardless of what the network is computing. In this case, a particular pair of function values (x_A, y_A) would belong to the activation level, while the function itself exists at the structural level[7]. This completes the

[7] Different interpretations of the node transfer function could lead to marginally different analyses in terms of the present framework. The exact identifications of model elements with analytic elements will typically not be nearly as important as consistency (once a determination for a class of correspondences is made, all others in that class should adhere to it) and careful bookkeeping to ensure no significant model element is identified with zero or more than one analytic elements.

translation of the node transfer function into the analytic framework I have proposed, as follows:

$$y_A = F_S(R_S, x_A)$$

All that remains is to analyze the perceptron at the network level. Plugging the node equation back into the network equation yields:

$$output_{Network} = y_i = vector_{all\_j}(F_S(R_S, x_A))$$

The network output (y_i) is clearly at the same level as the outputs from the individual nodes that comprise it:

$$y_A = vector_{all\_j}(F_S(R_S, x_A))$$

Finally, the vector(•) function is a specification of the network architecture. This function defines the activity-level function computed by the overall network, and will change depending on the structure of the task domain. In perceptrons, the network output is simply an aggregation of the individual node outputs in the form of a vector (hence, the function name "vector(•)"). For this reason, I have placed it at the activity level (F_A):

$y_A = F_A(F_S(R_S, x_A))$, where:
- x_A are the network inputs
- y_A are the network outputs
- R_S are the network weights and thresholds
- F_S = sgn(•) is the *basis function*
- F_A = vector(•) is a *network connectivity function* (or *composition function*)
- Note that M_{AS} is defined implicitly by the form of F_S and F_A

This completes the analysis of the perceptron equations. However, at the macroscopic level, many computational neuroscientists working with the perceptron were, like the cognitivists, trying to account for external behavior (Rosenblatt, 1958). That is, the measurable data used to define the inputs and outputs to the model were typically derived from behavioral observations, not measurements of internal activity

such as single-cell recordings, so these behavioral inputs (x_B) and outputs (y_B) have yet to be accounted for in this analysis. In terms of the framework in Figure 2.3, we need to specify the mapping from behavioral inputs to activity-level inputs (M^y_{BA}) and the mapping from activity-level outputs to behavioral outputs (M^y_{BA}). This requires a slight digression.

Shannon's Information Theory and the Information Equivalence Mapping

A powerful general principle used (at least implicitly) in virtually all theorizing and modeling is the idea of "information equivalence," mundane examples of which we all encounter regularly in our everyday lives, but which was not formalized until Shannon's (1948) paper A Mathematical Theory of Communication. The basic idea is simple enough to illustrate, as follows. Consider the numeric symbol "47", which specifies a well-defined referent (the number forty-seven). Given a set of discrete objects, for example, it is straightforward to determine unambiguously whether there are 47 of them or not. The abstract concept behind the number 47 does not change even if we give it a different label. For example, in binary notation (such as is used by a digital computer) the same number would be represented as 101111, and in Roman numerals it would be XLVII. Despite the differences in surface appearance, all three representations pick out the same entity (the number forty-seven) from the infinite array of possible referents (in this case, the natural numbers). We might say that these three representational systems exhibit *informational equivalence*, at least with respect to the natural numbers, because they cover the same field of reference, they pick out the same set of entities, and there is a reliable procedure for translating between them.

This general principle applies in a similar manner in the domain of modeling. When a cognitive scientist builds a production system model of problem solving behavior in the domain of balance beam problems, for example, she assumes that the features of the current balance beam problem encoded in the model's "working memory" (e.g., number of weights and distance from the fulcrum on each side of the balance) contain the same information as the features used by the human subject faced with the same problem.

Similarly, when an artificial neural network (like the perceptron) is applied to model behavioral data, the modeler assumes that the relevant information present in the external stimulus (the same balance beam problem, for instance) is preserved as it passes through the sensory system and appears as an activity pattern at the input to the network actually responsible for solving the problem. The format might change (even dramatically), but the assumption is that the core information is preserved.

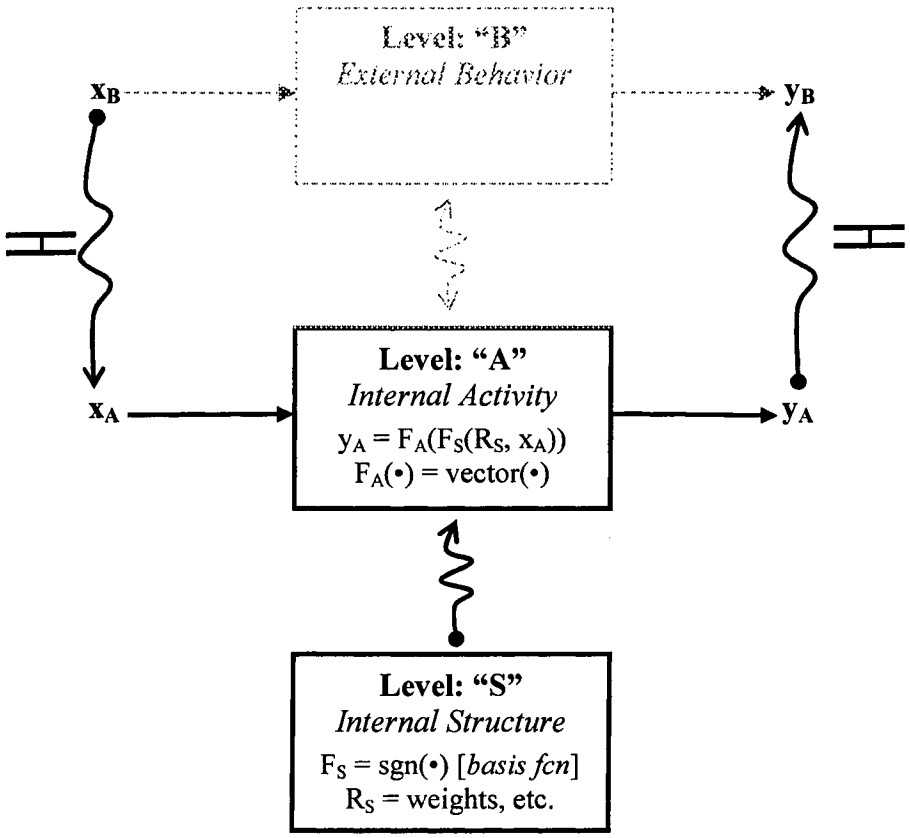
For present purposes, I have dubbed this the *information equivalence mapping*, and assigned it the symbol $\equiv$, which is a cross between a capital letter "I" (for information) and an equal sign (for equivalence). This, I submit, is the mapping (M^{x}_{BA}) that links the external behavioral inputs with the activity-level inputs of the perceptron and the mapping (M^{y}_{BA}) from the activity-level outputs of the perceptron to the behavior-level output being modeled by the system:

$$x_A \equiv x_B$$
$$y_B \equiv y_A$$

The complete perceptron analysis is summarized graphically in Figure 2.9.

In an interesting historical parallel, proponents of the symbolic paradigm used methods similar to those they leveled against behaviorism (for example, proof-by-counterexample) to eliminate the perceptron as a viable basis for a theory of psychology

Figure 2.9: A perceptron model represented in terms of my analytic framework

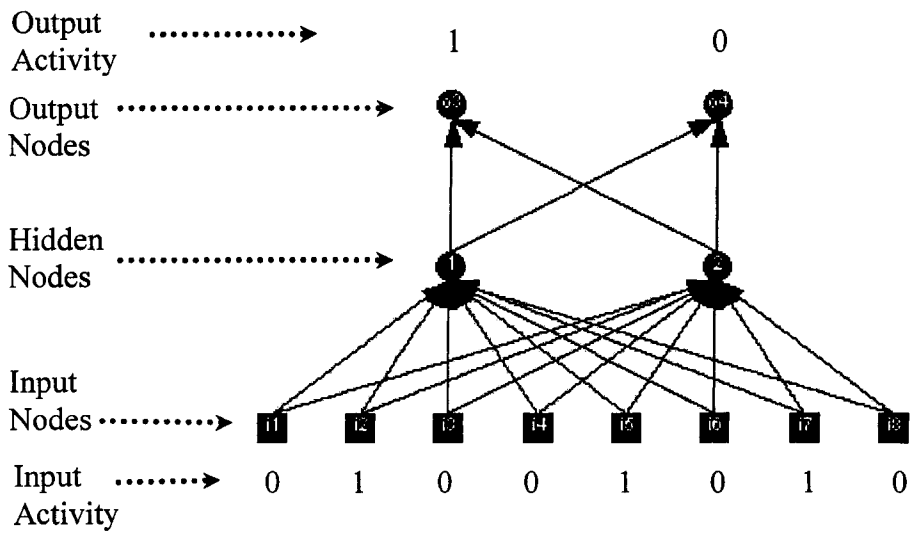


(Minsky & Papert, 1969). In effect, Minsky and Papert demonstrated that the perceptron could not—even in principle—model certain capabilities that people routinely demonstrate (for example, judging whether a geometric figure is closed or not). The critical problem derives from the form of F_A (the network connectivity function); with only a single layer of nodes, the network is only as powerful as any single element in it, and a perceptron node can only solve a relatively simple class of problems that are *linearly separable* (see Minsky & Papert, 1969 for a careful treatment of these issues; see Elman et al., 1996 for a less technical but more readable account). Researchers knew that the limitations of the perceptron could be overcome by adding an additional layer of nodes and connections between the inputs and outputs. However, it took more than fifteen years of additional theoretical and applied work combined with the advent of fast and inexpensive computers to demonstrate the capabilities of the refined models and generate substantial interest in them among researchers in the field (Anderson & Rosenfeld, 1998).

Computational Neuroscience: Multi-Layer Perceptrons

Minsky and Papert all but extinguished ANN research when they so decisively eliminated the perceptron as a framework for modeling human cognition (Anderson & Rosenfeld, 1998; Gardner, 1985). Researchers eventually figured out how to overcome the technical challenges involved in adding additional layers of nodes with modifiable connections to the basic perceptron (Anderson & Rosenfeld, 1998; Gardner, 1985; Rumelhart et al., 1986), introducing what is sometimes called the "multi-layer perceptron" (or MLP for short; see Figure 2.10) to distinguish it from the original single-layer perceptron (Figure 2.8).

Figure 2.10: Example of a multi-layer perceptron with two layers of modifiable connections between input and output nodes



Mathematical Definition of the Multi-Layer Perceptron

Like the original perceptron, the MLP is intended by many to model the behavior of neurons or collections of neurons in the nervous system (recall Figure 2.7; Elman et al., 1996; McClelland & Rumelhart, 1986; McLeod, Plunkett, & Rolls, 1998; O'Reilly, 1999; O'Reilly & Munakata, 2000; Rolls & Treves, 1998; Rumelhart & McClelland, 1986).

For a network of units working together on a single problem, the j^{th} node computes a sigmoid function of its inputs, as follows[8]:

$$output_j = 1/(1 + \exp-(\Sigma_i\ weight_i * input_i))$$

The function on the right-hand side is called a "sigmoid" function (it looks like a squashed-S). Simplifying the notation, therefore:

$$output = sigmoid(weights, inputs)$$

The analysis of this equation is essentially the same as for the perceptron node transfer function in the previous section. The result is the same generic equation:

$$y_A = F_S(R_S, x_A)$$

Where:

$F_S(\bullet)$ = sigmoid($\bullet$) is the basis function for this MLP instead of sgn($\bullet$)

At the network level, the connectivity function is different for the MLP than it was for the perceptron (most importantly, as I mentioned, because the MLP has more than one layer of modifiable connections). Instead of simply aggregating the node outputs into a vector and taking that as the network output, in the MLP the function relating individual nodes to overall network function (which I am calling network($\bullet$)) is

[8] There are actually a variety of functions used for node transfer functions in MLPs by different researchers. They should all lead to the same end result when analyzed in this framework. This is one of the benefits of this analytic framework—it allows for the suppression of a lot of variation in the model details in order to highlight the key definitive differences between model families and theoretical paradigms.

more complicated in its details but again results in an equation similar to that of the perceptron:

$y_A = F_A(F_S(R_S, x_A))$, where:
x_A are the network inputs
y_A are the network outputs
R_S are the network weights and thresholds
F_S = sigmoid(•) is the *basis function* for this MLP
F_A = network(•) is a *network connectivity function* specific to each MLP

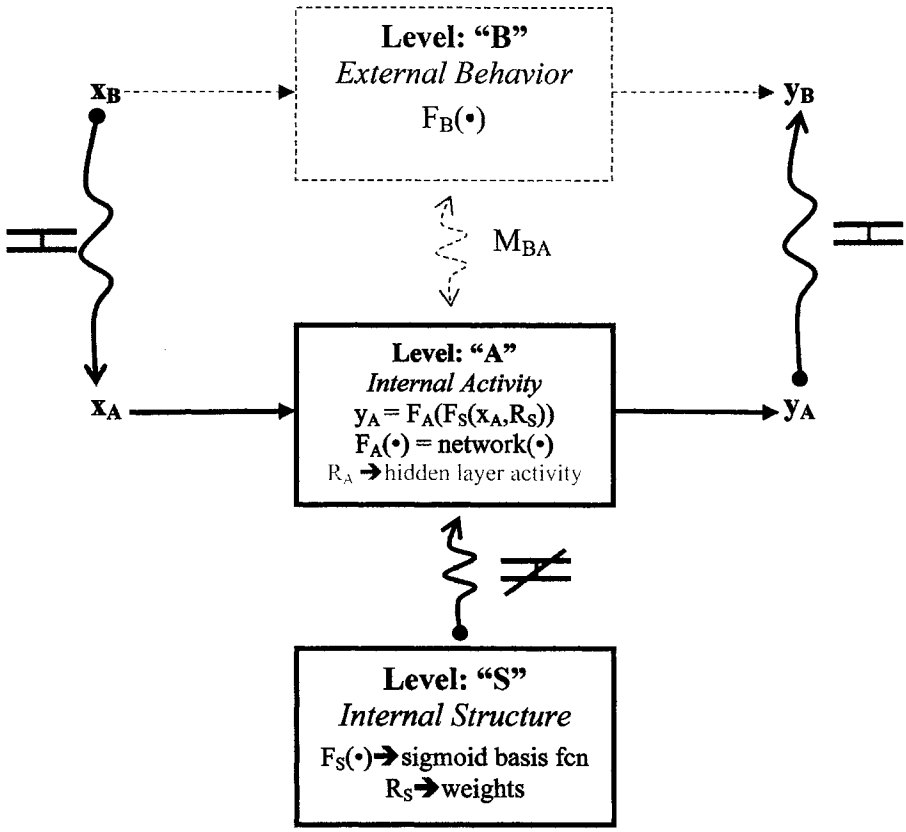
In addition, the same reasoning applies as for the perceptron in the following cases:

$M^x_{BA} =$  (information equivalence mapping)
$M^y_{BA} =$ 

The results of the analysis are summarized graphically in Figure 2.11.

So far, this analysis foregrounds the parallels between the original single-layer perceptron (Figure 2.9) and the MLP (Figure 2.11) at the abstract level. The differences are mostly abstracted away in the details of F_S (the basis function) and F_A (the architecture specification or connectivity function). There is a difference arising from the additional layer(s) of the MLP, however, that needs to be accounted for explicitly here. The MLP introduces a layer of "hidden nodes" whose activity-level outputs are not accounted for in either the network inputs (x_A) or outputs (y_A). These patterns clearly belong at the internal activity level, and they are the products of functions (not functions themselves), so they can only be identified with R_A–that is, activity level representations (see Figure 2.11). Indeed, a number of researchers have pointed to these hidden layer activities as providing a neurologically grounded model of our cognitive internal representations (Elman et al., 1996; Laakso & Cottrell, 2000; McLeod et al., 1998; Quinn & Johnson, 1997).

Figure 2.11: A multi-layer perceptron (MLP) model represented in terms of my analytic framework



Note that M_{AS} is defined implicitly by the forms of F_S and F_A. One interesting property of this mapping is that, unlike the other two (M^x_{BA} and M^y_{BA}), it does not exhibit information equivalence ($\not\equiv$). Specifically, the information encoded in the structural representations (R_S=weights, thresholds) is not equivalent to the information encoded in the activity-level internal representations (R_A=hidden layer outputs) or the activity level inputs and outputs (x_A and y_A).

When people talk about "distributed representations" in neural networks, they typically fail to specify whether they mean the distributed structural representations (R_S=weights) or the distributed activity-level representations (R_A=activation patterns). The distinction is important, for the reason just cited: *although the structural representations and activity-level representations are both distributed, these two sets of representations are not informationally equivalent*[9] (that is, they contain different information). I submit that it is this feature of neural networks, not the existence of distributed representations *per se*, which differentiates ANNs from other types of computational models (especially those based on the symbolic paradigm). The fact that ANNs (at least when offered as models of biology) are parallel (not serial), use distributed representations (not localist), are sub-symbolic (not symbolic), exhibit graceful degradation (not catastrophic failure), and support content-addressable memory are all direct consequences of this more fundamental fact, not the other way around. I argue later in this paper that this insight can be used to differentiate between the symbolic

[9] For example, I just saved the document I am editing in my word processor to the hard disk. The version of the document that I am editing in the computer's working memory (RAM) and the version just saved to hard disk use different representational media and formats, but they contain the same information (words, formatting commands, etc.)—they are informationally equivalent. In other words, at an abstract level they are basically copies of one another even though they exhibit superficial differences. Earlier today I took a picture of my son using my digital camera and copied the photo onto this computer's hard disk. This electronic manuscript and that photograph are obviously not copies of one another at any level—they contain very different information. Therefore, those two files are not informationally equivalent.

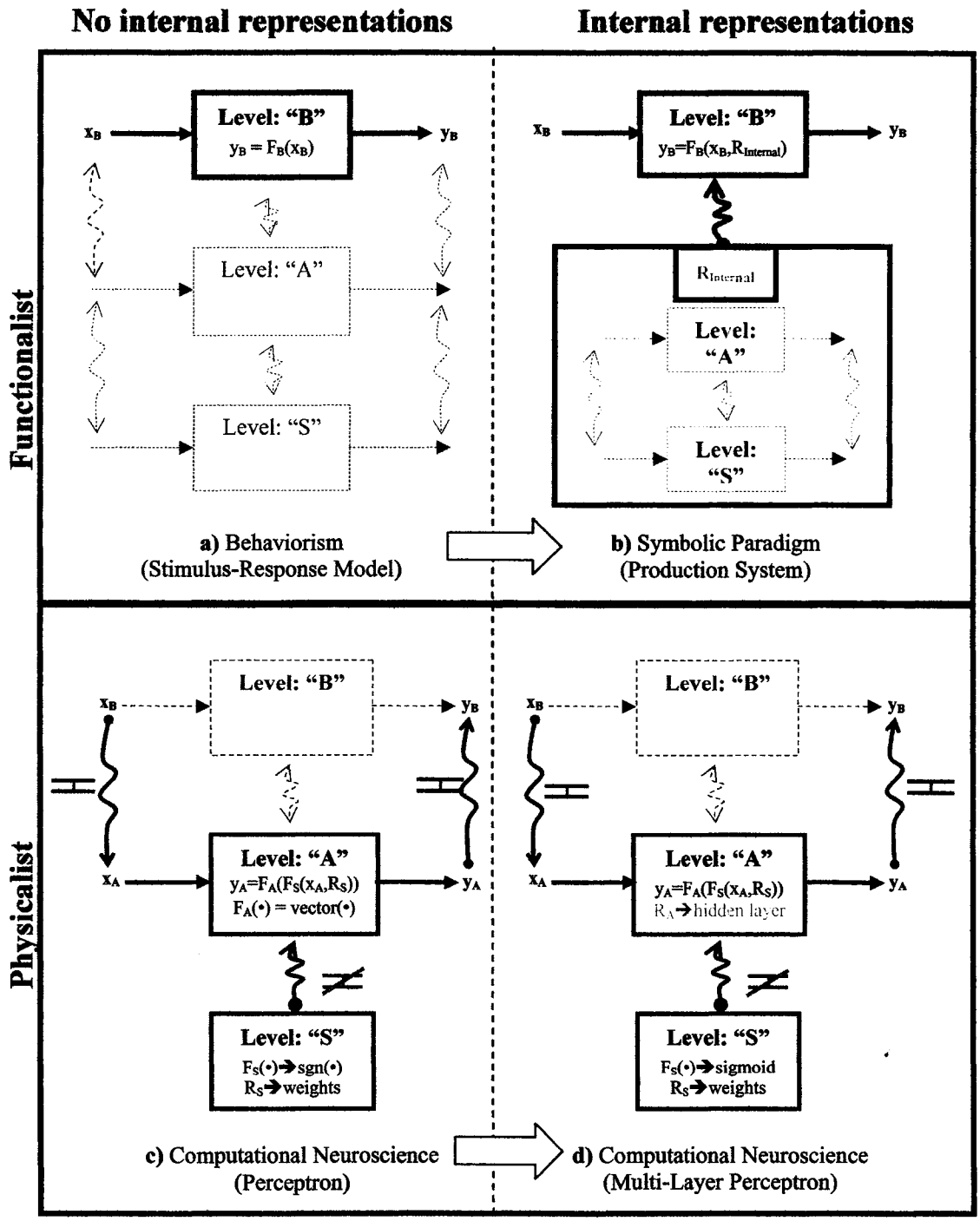
paradigm and the ANN paradigm (although not in the obvious way through direct comparison). I discuss these issues further in a later section.

Insights from the Intertheoretic Analysis

Figure 2.12 summarizes the results of the four analyses. The similarities and differences between the frameworks emerging from my analysis are neatly summarized in a two-by-two table. Along the vertical axis, the behaviorist model and the production system are grouped together under the "functionalist" heading while the perceptron and the MLP are grouped together under the "physicalist" heading to reflect their different primary foci on external and internal data, respectively. On the horizontal axis, the behaviorist model and the perceptron are categorized as having no internal representations, while the production system and MLP are both identified as supporting internal representations.

In an interesting historical parallel, the fundamental problem with the perceptron is quite similar to the fundamental problem with behaviorism, as this diagram makes clear: both lacked a mechanism for representing information internally, which made them demonstrably too weak to model significant features of human cognition and behavior. Proponents of the symbolic paradigm used similar methods to provide the decisive blows to both frameworks, which was undoubtedly a major factor in the subsequent ascendancy of that framework for several decades. As I have described, the cognitivist response to the behaviorist problem was basically to add the capacity for internal representations to the behaviorist framework (compare the top two panels in Figure 2.12). Similarly, the response of computational neuroscientists was to add what is called a "hidden layer" to

Figure 2.12: Taxonomic summary of the analyses of four psychological models: a) behaviorist stimulus-response model, b) symbolic paradigm production system model, c) single-layer perceptron, and d) multi-layer perceptron (MLP).



the perceptron that comes between the input and output layer, providing the capacity for internal representations in that model (compare the bottom two panels in Figure 2.12).

In both cases, the change to the paradigm looks on the surface to be quite superficial from its predecessor. The consequences, however, are profound. Whereas the original frameworks (behaviorism and the perceptron) were inadequate to model human cognition, the paradigms that replaced them (the symbolic paradigm and the MLP, respectively) are provably able to model *any* set of input-output transformations.

The major difference between these two historical progressions can be seen along the vertical axis in Figure 2.12: the way the theories are constructed is quite different (and complementary) in terms of the material bases in which they are grounded. Behaviorism and the symbolic paradigm are grounded in behavioral data, and insist on maintaining a black box around the internal mechanisms of cognition, beyond the demonstrable need (in the case of the symbolic paradigm) to allow for internal representations. The perceptron and MLP, in contrast, are derived from observations about the structure and function of the nervous system (see, for example, the derivation of a simplified neural network model from a detailed biophysical neural model in Ermentrout, 1994), which are then used to investigate when and how these kinds of internal structures and processes can give rise to observable patterns of behavior.

The comparison in the bottom row demonstrates why being biologically based does not in itself make for a better or more plausible model. The perceptron is implausible and the MLP is plausible, even though both are quite similar in terms of their biological grounding.

This analysis also suggests how one might construct a typology for systematically organizing extant theories (Figure 2.12 is one example of how that might look), and perhaps a tool for identifying new possibilities that have not been previously explored. For example, none of the frameworks considered in the present analysis included any representations at the behavioral or external level (R_B). Theoretical paradigms that might need to incorporate such representations include the contextualist (Rogoff, 1990) and the situated cognition frameworks (Lave & Wenger, 1991), both of which put a heavy emphasis on the role of external symbols and artifacts in cognition and behavior.

The present analytical framework should be able to accommodate such external representations, and the foregoing analysis highlights two sets of issues that would arise in connection with their use. First, the current consensus among computer scientists is that a theoretical framework that incorporates external representations cannot in principle be any more powerful than either the production system or the ANN modeling frameworks (Chown, 2004). It is not immediately obvious, therefore, what additional theoretical leverage is gained by appealing to external representations beyond what is afforded already by internal representations. Moreover, an account of the leverage gained from external representations cannot be based on the assumption that these other models are somehow representationally impoverished. Although any given production system or ANN might exclude critical explanatory variables, in principle the missing variables could be accommodated by the general modeling frameworks from which these particular models are derived. While it is very possible that symbolic and neural modelers could benefit from the shift of emphasis or broadened view of the contextualist and/or

situated cognition perspectives, it is not obvious on the face of it that the underlying theoretical frameworks offer any additional explanatory power in principle.

Second, an account of the relationship between external representations (for example, the words printed in a book) and internal representations (for example, the contents of working memory while the book is being read, or the changes to long term memory resulting from studying the book) would be necessary to avoid the critical paradigmatic problems faced by the behaviorists. Although it is clear that externalized cultural structures, practices, and artifacts have a central role in shaping behavior, the effects of such external representations are also clearly mediated in every case by the neural and cognitive structures of individual persons. In this sense, the external representations are redundant with the internal representations that are generated when people interact with them, and the internal representations arguably have greater epistemological primacy since they are the structures implicated in carrying out the actual work. In addition, the diverse interpretations that can be imposed on a cultural artifact like a book arise not from the book itself—which is largely invariant from one encounter to the next—but from the interaction of diverse brains and minds with the artifact. For example, the evolution of human languages is driven in large part by the interaction between external representations (e.g., spoken and written words) and variation in the internal representations of individuals who use a language and teach it to others (Tallerman, 2005). Consequently, internal representations must be included (either explicitly or implicitly) in any theory of external representations that acknowledges the kind of representational variability just described, but the converse is not necessarily true.

Taking a step back, it appears that the cognitivists and the computational neuroscientists are not arguing so much about how the world *is* as about how the world *can best be described.* Evidently, a direct functionlist vs. physicalist comparison is not going to cut very deeply because both are equally powerful and both are fundamentally materialist; they are just grounded in different material information sources. In particular, the kinds of tactics used by the cognitivists to discredit the two frameworks in the left-hand column in Figure 2.12 will not be able to differentiate between those on the right. Moreover, the two frameworks on the right are as powerful as any currently conceivable computational systems can be (Haykin, 1999), which means that purely analytic strategies are unlikely to be effective in resolving the longstanding debate over which is the superior framework for modeling human cognition (Chown, 2004). It seems to me that a new and distinct approach is called for in order to differentiate between these two frameworks, one perhaps based on empirical methods instead of analytic. In the next section I explore some possibilities in this area, by applying results of the previous analysis to identify promising entry points.

Previously Proposed Bases for Comparing the Symbolic Paradigm to the ANN

Many attempts have been made (mostly by ANN researchers) to distinguish between the symbolic paradigm (e.g., production systems) and the MLP (or other multi-layer ANNs) on various bases. For example, computational neuroscientists point out that ANNs tend to use distributed representations while production systems tend to use localist or modular ones (Elman et al., 1996; McLeod et al., 1998), that ANNs are parallel and production systems are much more serial (Elman et al., 1996; McLeod et al., 1998), and that ANNs operate on "subsymbolic" data structures while production systems

operate on "symbolic" ones (Smolensky, 1988). ANN researchers also often point to intrinsic characteristic behaviors of ANNs that mirror characteristic behaviors of biological nervous systems, including graceful degradation, spontaneous generalization, and content-addressable memory (Elman et al., 1996; McLeod et al., 1998; Smolensky, 1988), which are not shared intrinsically by production systems.

In my view, none of these attempts to differentiate ANNs from production system models is very convincing. There are two sets of responses from the symbolic camp to each of the proposed dichotomies, one particular and the other general. The particular response in each case basically points out why the difference alluded to is not fundamental (Klahr & MacWhinney, 1998). For example, one can prove using computer science theorems that any parallel algorithm can be converted into a serial algorithm that performs the same set of computations (but not the other way around—Marr, 1982). In addition, the distinction between "distributed" representations and "localist" representations is not well-defined (Klahr & MacWhinney, 1998). The text of this paper is spatially distributed across each page and also across many pages, yet this would be considered a paradigmatically localist representation. Similarly, a "symbolic" model could be made progressively more "sub-symbolic" if the granularity of its inputs, representations, and outputs were made finer and finer.

The more general categorical refutation of all such attempts to differentiate ANNs from production systems is based on the fact that these are all accidental—not essential—properties of a particular production system implementation. In other words, these are by and large all physicalist details of the model to which a functionalist is not theoretically

committed one way or another (Churchland, 1988; Fodor, 1968; Harman, 1989; Levin, 2004; Sellars, 1954).

If none of these arguments is conclusive, then why do they persist? The intertheoretic analysis conducted above gives insight into why symbolic models like production systems and artificial neural networks like MLPs are not directly comparable, why the debaters seem to be talking past one another so often, and therefore why the debate persists without resolution.

First, if we were to overlay the top and bottom panels in the right column of Figure 2.12, we would find that the only immediate correlation involves the external inputs (x_B) and outputs (y_B). The other behavior-level elements of the symbolic paradigm (F_B) have no counterpart in the MLP. This mismatch reflects the fundamental difference between the functionalist and physicalist accounts of a phenomenon or task domain (for instance, balance beam problem solving)—in this case it is a comparison between functionalist apples (F_B) and physicalist oranges (F_A, etc.).

The remaining elements in the two diagrams do overlap—and must therefore refer to the same ontological referents. The difference has to do with their degree of referential transparency. The computational neuroscientists, by and large, are attempting to specify how the elements of the model map onto internal structures and processes—their ultimate goal is total referential transparency. The symbolic paradigm researchers, in contrast, acknowledge the theoretical necessity of internal representations, but insist—even in principle—on complete black box opacity regarding the internal structures and processes supporting those representations. The incomparability here arises from the attempt to compare the computational neuroscientist's individual apples (e.g., F_A, R_A) and oranges

(e.g., F_S, R_S) to the symbolist's bag of mixed fruit ($R_{Internal}$). This is how the two frameworks can ostensibly refer to some of the same ontological entities (internal representations) without being directly comparable on any single point.

So, how can these two frameworks be compared? The most obvious strategy follows from the preceding observations, and involves comparing them on the only elements that both overlap and unambiguously share common referents: the external inputs (x_B) and outputs (y_B). In fact, this is what is often done. Given a single set of observations in a well-specified task domain (such as the balance beam task), researchers who want to evaluate the performance of a particular model or compare the two frameworks construct (or otherwise acquire) symbolic models (e.g. production systems) and/or artificial neural network models (e.g., MLPs) of the target behavior (solving balance beam tasks). They then evaluate and/or compare the models based on their relative goodness-of-fit to the observed data (see, for example, Besner, Twilley, McCann, & Seergobin, 1990; Elman et al., 1996; Klahr & Siegler, 1978; McClelland, 1989; McClelland & Jenkins, 1991; McLeod et al., 1998; Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989; Siegler, 1976, 1981; Stoianov, Stowe, & Nerbonne, 1999). This approach has some utility, but it has a significant drawback.

As I mentioned earlier, production systems and MLPs are equally powerful, and indeed both are powerful enough (given sufficient time, representational resources, and programming finesse) to model *any* input-output behavior. So when one model performs better than another, the other camp simply has to add more memory or nodes and perhaps tinker with the programming a bit until it matches or outperforms its rival, and this is often what occurs in practice (in the case of regular and irregular verb processing

models, see Besner et al., 1990; Coltheart, 1978, 1985; Coltheart, Curtis, Atkins, & Haller, 1993; in the case of the balance beam models, see Inhelder & Piaget, 1958; Klahr & Siegler, 1978; McClelland, 1989; McClelland & Jenkins, 1991; McLeod et al., 1998; Plaut et al., 1996; Seidenberg & McClelland, 1989; Siegler, 1976; Stoianov et al., 1999). This game of leapfrog can continue *ad infinitum* without any real hope of ultimately resolving the issue (Bechtel & Abrahamsen, 1991)—certainly not without additional theoretical constraints (some candidates for which I discuss in a later section of this paper).

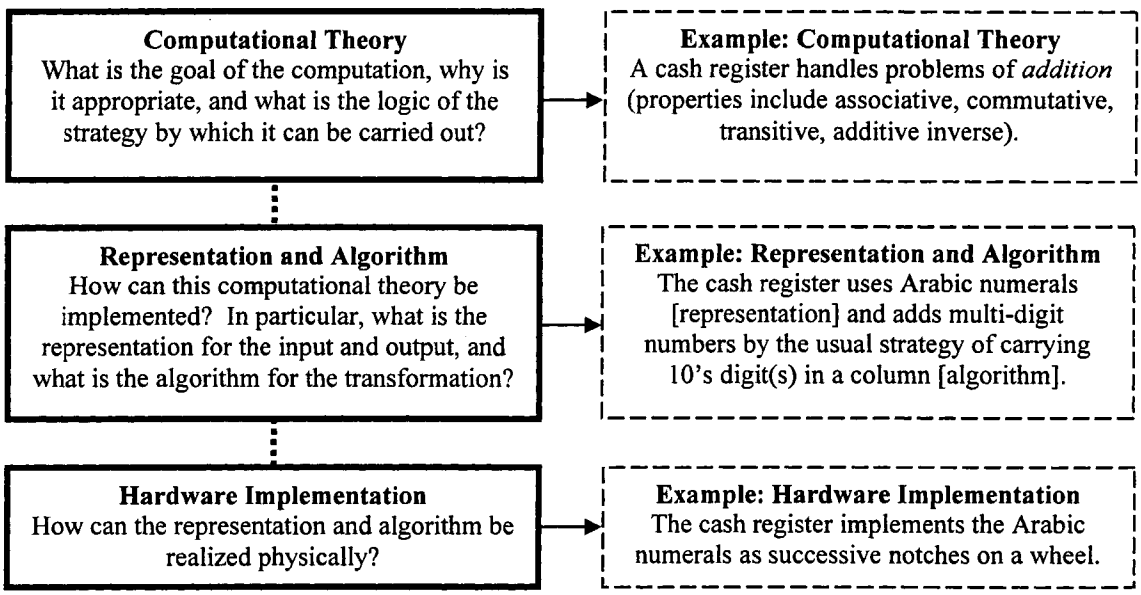
Part of the problem, I would argue, is that it is not widely recognized why (or perhaps even that) these frameworks are not comparable, because the predominant analytic framework within which these issues are usually discussed (that is, Marr's "levels of explanation"—see MacDonald & MacDonald, 1995) entails ambiguities that obscure this important point. In the next section, I apply my analytic framework and the results derived in the previous sections to try and pinpoint the source of this obscurity and suggest a way to resolve it.

A Comparison of the Production System and the MLP

Marr's Three Levels of Explanation

Marr (1982) proposed that a full understanding of any information processing system (e.g., the human visual system) ultimately involves understanding it from three perspectives (Figure 2.13): computational theory (computation), representation and algorithm (algorithm), and hardware implementation (implementation). Marr's levels of explanation are widely accepted and employed by cognitive scientists (MacDonald &

Figure 2.13: Marr's three levels of analysis (left-hand side) and his cash register example (right-hand-side)



MacDonald, 1995; Posner, 1989). Marr (1982) offers the example of a cash register to make the definitions concrete (Figure 2.13).

The *computation* level is an abstract specification of the task domain; in other words, it specifies the inputs and outputs and describes the computational goal of the system, along with the rationale for carrying out the computation in the context of the task (also sometimes called "semantics"—Posner, 1989). A cash register handles addition problems, which is appropriate when people are making purchases because, for example, buying nothing should cost nothing (additive inverse); buying several items together should cost the same as buying them separately (associative property of addition); and purchasing the same set of items at fixed prices should always result in the same total cost, regardless of the order in which they are purchased (commutative property of addition). In terms of my framework it can be seen that this "level of explanation" is really a formal specification of the information equivalence mapping obtaining between objects (e.g., prices of consumer items) and operations (e.g., tallying purchase totals) in the world and elements (representations and transformation functions) in this particular model.

The *representation and algorithm* level specifies a format for the information involved (e.g., Arabic numerals) and an effective procedure that can carry out the computation on that representation (for example, the familiar method of carrying tens). Some people have called this the "syntax" specification for the system (Posner, 1989).

Finally, the *implementation* level specifies the details involved in realizing the computation in hardware (e.g., in terms of digital bits, transistors, synapses, spike trains, etc.). Marr considered this level to be probably beyond the purview of cognitive science.

There is some ambiguity in the level definitions (Posner, 1989), but it seems clear enough that a production system model belongs at the "representation and algorithm" level, since it provides a representation for the input and output (encoded in "working memory") and an effective procedure (or algorithm) for transforming inputs into outputs (comprised of the set of productions).

Although it might appear that an MLP neural network could belong at Marr's implementation level, this is not the case—MLPs could be implemented either on digital computers (as ANNs typically are, as a matter of fact) or in a network of biological neurons (as a biological neural network). Furthermore, the MLP, like the production system, specifies a set of inputs and outputs along with an effective procedure for transforming inputs to outputs. Therefore, production systems and MLPs both belong at Marr's "representation and algorithm" level of explanation.

A conflict arises at this point. Production systems and MLPs are on the one hand specified in terms inviting direct comparison: both are effective procedures dealing in many cases with the same external inputs and outputs that can be simulated on a computer, and both belong at Marr's representation and algorithm level of explanation. On the other hand, I argued based on my analysis that despite appearances they are not really directly comparable at all (recall Figure 2.12b and 2.12d).

The resolution of this dilemma comes from recognizing that Marr's "representation and algorithm" level allows for (at least) two distinct theoretical renderings of any given model: functionalist and physicalist. For example, the production system analysis earlier was based on the functionalist view, with the result shown in Figure 2.12b. This result reflects the fact that the functionalists do not want to

make explicit commitments concerning the ontological referents of the production system model's internal mechanisms.

If we were to take the stance (as a thought experiment) that the elements of the production system model could be correlated with specific entities in the nervous system, we would end up with a different result. Assume for the moment that humans have a working memory that is structured like a production system's. A production system working memory is typically based on a computer's *random access memory* (RAM) which is a vast array of binary switches called *bits* that can take the value of 0 or 1. Working memory in such a phsyicalist production system would be most analogous to the inputs, outputs, and internal representations in an MLP, so these elements would in this case be identified with activity-level inputs, outputs, and representations (x_A, y_A, and R_A, respectively). Production memory is also typically based on RAM. The productions in production memory are most analogous to the functions computed at the network level in an MLP (F_A). The basis functions and durable representations modifiable via learning belong at the structural level. In a computer, the analog of basis functions would be the intrinsic operations of the central processing unit (identified with F_S). The durable, modifiable representations (R_S) are those stored on devices like hard disks.

In summary[10]:

$y_A = F_A(x_A, R_A)$, where:

y_A are the activity-level outputs (e.g., the solution stored in working memory)
x_A are the activity-level inputs (e.g., the initial problem encoded in working memory)
F_A are the production rules

[10] I have made an effort to make this case consistent with the MLP analysis, which also facilitates a comparison of the two results. Some readers might object to the particular details of certain identifications I have made. For example, it would be possible to reserve part of RAM to store the results of learning rather than saving them to a hard disk. I have made these examples concrete to clarify the exposition, and nothing important to the analysis hinges on these particulars.

R_A are the contents of working memory other than inputs and outputs (which could be the results of previous computations as well as information read into working memory from long-term memory (R_S))
R_S are the contents of a hard drive or other medium that stores the results of learning long-term (that is, beyond the duration of the current task)
F_S are the primitive operations of the computer's central processing unit

In addition, the same reasoning would apply as for the MLP in the following cases:

$M^x_{BA} =$ ⊣⊢ (information equivalence mapping)
$M^y_{BA} =$ ⊣⊢

These results are summarized graphically in Figure 2.14.

In some symbolic models, information must be retrieved from the more durable structural representations (R_S) in the course of solving a problem (for example, imagine the constant *pi* has been learned by the system earlier and is needed currently to calculate the circumference of a circle). Before this information can be used, it must be loaded into working memory (R_A). The most common way to implement this in a symbolic system would be to read the pattern of bits representing the constant *pi* directly from the disk into RAM. In other words, the information stored in R_S and R_A would be informationally equivalent in this typical scenario (R_S ⊣⊢ R_A). This mapping can be identified with M_{AS} on the generic analysis diagram (Figure 2.3).

The results of this (hypothetical) physicalist analysis of the production system are shown in Figure 2.15 (bottom right), alongside the physicalist analysis of the MLP (Figure 2.15, bottom left), and below the functionalist analysis of the production system (Figure 2.15, top). This figure highlights a number of important insights.

Earlier I argued that the functionalist production system is not directly comparable to the physicalist MLP. This diagram makes it clear why that is true—because there is another interpretation of the production system model that *is* directly

Figure 2.14: The result of a thought experiment, showing how a production system model would be represented in terms of my analytic framework, assuming the theoretical stance that elements in the production system model should be interpreted as representing physical entities in the nervous system

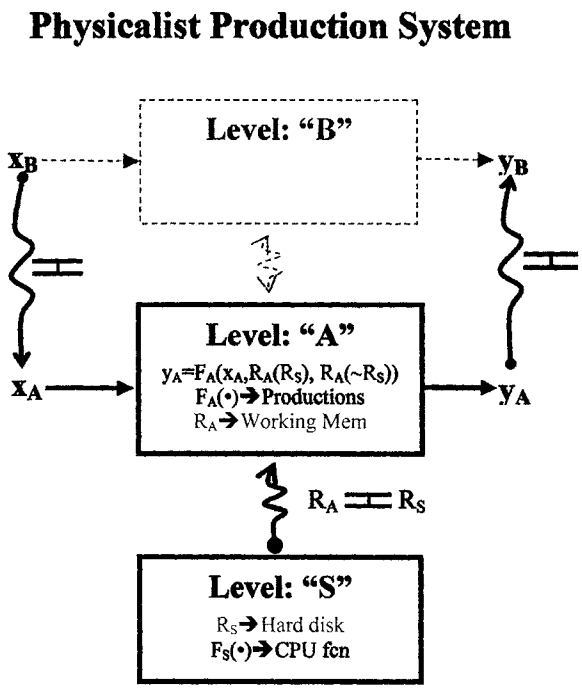
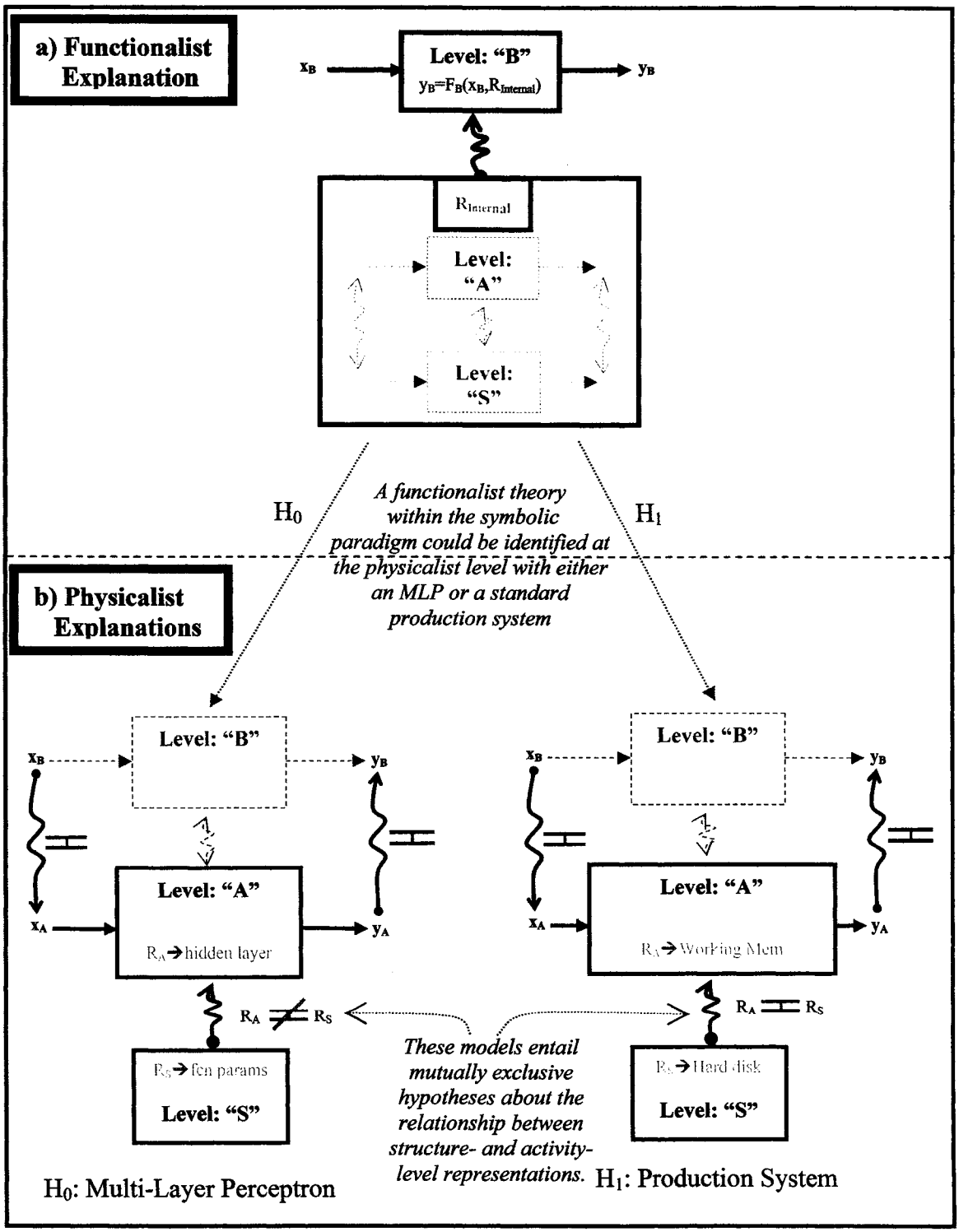


Figure 2.15: The functionalist production system (a) does not make theoretical commitments concerning the referential relationship between its internal representations and entities in the nervous system, so it could in principle be implemented as (b) either a physicalist MLP or a physicalist production system

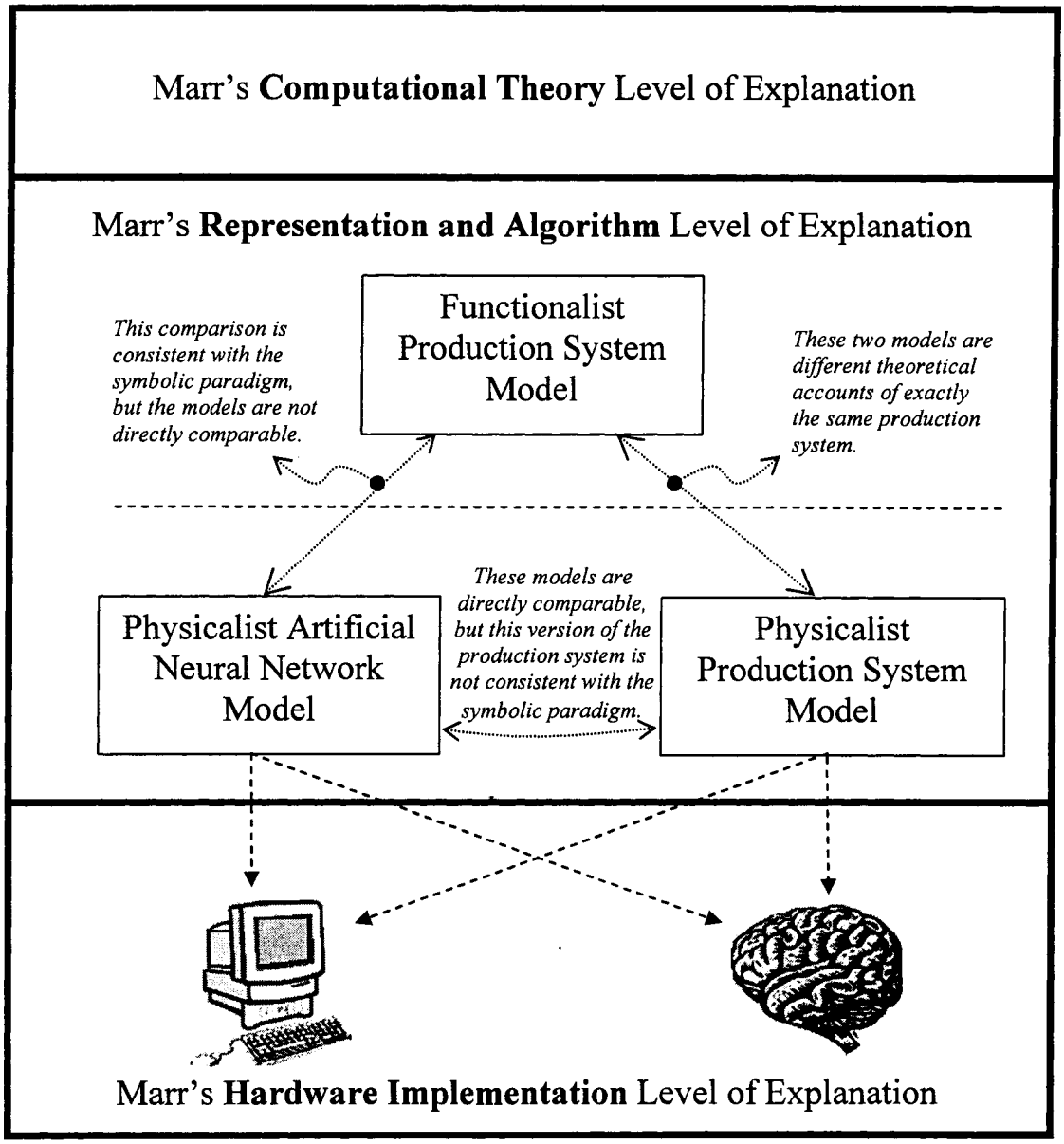


comparable to the physicalist MLP. The production system on the bottom is comparable to the MLP, but no mainstream proponent of the symbolic paradigm holds that view of the production system model (Klahr & MacWhinney, 1998; Pinker, 1997)—the official party line is the view at the top. Moreover, the functionalist production system model (at the top) could in principle be realized using either a physicalist production system or a physicalist MLP (and there might be other options). Note, again, that these two models are not implementation models in Marr's scheme, because each of them, in turn, could be realized in either a digital computer or a biological neural network (see Figure 2.16 for a simpler graphical summary of this point).

As I discussed, the two physicalist models entail logically exclusive hypotheses about the relationship between structural representations (R_S) and activity-based representations (R_A). This physicalist interpretation of the production system model embodies the hypothesis that at least some of the structural representations (e.g., files stored on hard drives) are informationally equivalent to some of the activity-level representations (e.g., the contents of those same files loaded into RAM for processing).

The MLP model embodies the alternate hypothesis that the structural representations (e.g., connection weights and node thresholds) and activity-based representations (e.g., network activation patterns) are not informationally equivalent. Although both hypotheses can be true of different parts of the same system at the same time, they cannot both simultaneously be true of the same exact referent. I propose that these two mutually exclusive hypotheses could serve as the basis for empirical experiments to differentiate between different physicalist models (I describe one such experiment in chapter 3).

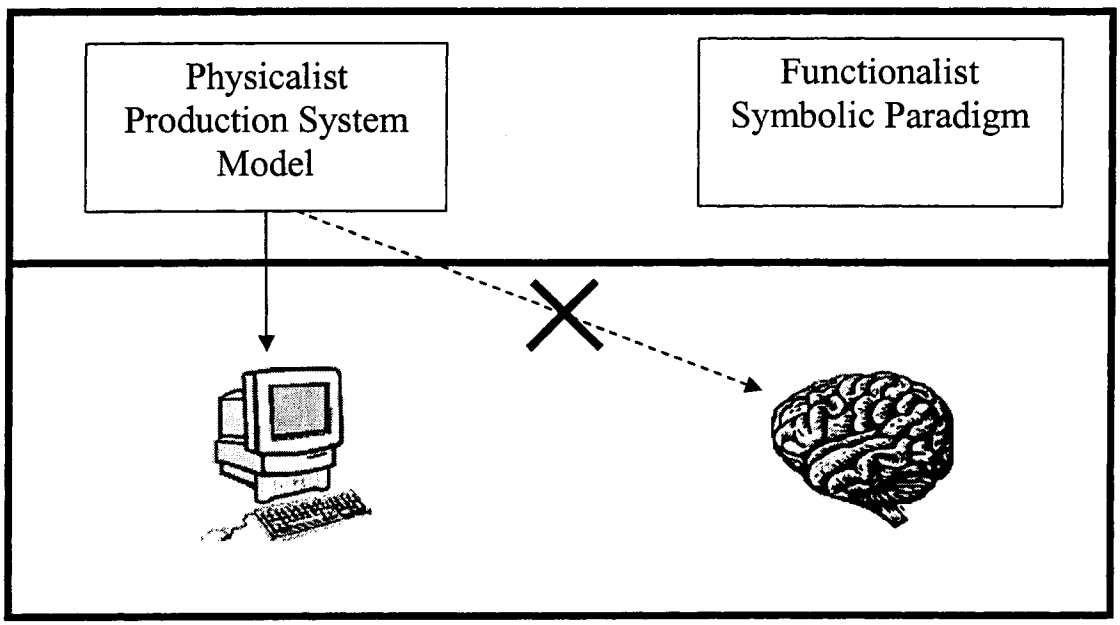
Figure 2.16: A problem with Marr's levels of explanation is that there is one missing



Furthermore, the fact that the functionalist model (at top) accommodates both of these mutually exclusive hypotheses suggests that it is not falsifiable (at least with respect to this feature), and therefore not open to scientific investigation. The MLP, in contrast, makes a commitment to a particular kind of mechanism, which exposes it to empirical/scientific study in ways not possible for the functionalist symbolic paradigm. I propose that this is how the MLP can be differentiated analytically from the production system as it is conceived by the mainstream symbolic camp. In the next chapter I describe the results of an experiment designed to test aspects of the MLP hypothesis in this regard.

Both of the production system diagrams (functionalist and physicalist) are derived from the same physical production system model, and both belong at Marr's "representation and algorithm" level of explanation. This example highlights a source of serious ambiguity in Marr's formulation, since these two theoretical interpretations of the same physical model have very different consequences, as I have described. That is, Marr's "representation and algorithm" level of explanation can accommodate very different types of descriptions (what I am calling "functionalist" and "physicalist" models), and as a result the important differences between these types of models cannot be distinguished within his framework. This ambiguity is probably also one reason why the connectionists and symbolists often seem to be talking past one another, because they really are—the symbolists are arguing *from* a *functionalist* symbolic paradigm perspective while the connectionists are arguing *against* a *physicalist* production system perspective (see Figure 2.17)—but this discrepancy is not apparent from within Marr's original framework.

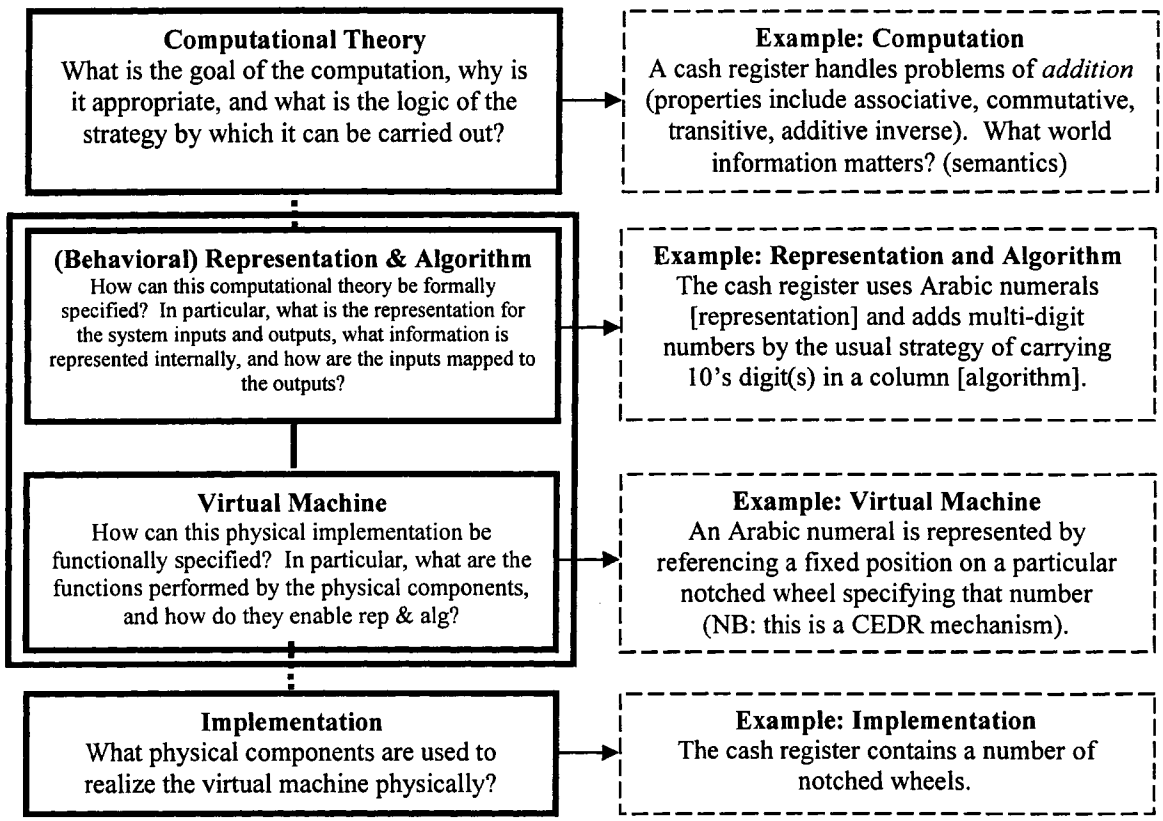
Figure 2.17: Confusion arises when researchers mistakenly assume a physicalist production system model is intended as a claim about neural implementation



Based on the present analysis, I argue that Marr is missing a necessary level of explanation. We might call this the "virtual machine" level of explanation and place it between the current "representation and algorithm" level and the "mechanism" level (Figure 2.18). Alternatively, we might divide the "representation and algorithm" level into two new levels: functionalist-representation-and-algorithm and physicalist-representation-and-algorithm. Explanations would be placed at one or the other level depending on their level of referential transparency, as discussed above. This deeper difference between theories has perhaps been further obscured by the oft-repeated observation that two representations with different surface formats can nonetheless contain the same information. We can all agree that surface format does not matter—but that does not imply (as this argument demonstrates) that there are no meaningful differences in representational types once we have controlled for surface format (see Bishop, 1995 for a technical overview of different classes of representational systems).

Marr's cash register example can now be seen as an unfortunate choice to illustrate his levels of explanation, for two reasons. First, the cash register is basically a behavioristic system in the sense that it requires no internal representations to perform its function, because it can simply output a partial sum after each step and then take that partial sum as one if its inputs for the next step. (For this reason it is also a rather peculiar example to use, given the centrality of internal representations to cognitive science's world view.) In a system like this where no internal representations are needed, there is really no pressing need to separate the virtual machine level of analysis from the implementation and/or representational levels because there is little internal complexity to model beyond what is specified in the representation and algorithm. In a system with

Figure 2.18: A revised version of Marr's levels of explanation, illustrated with a revised example



internal representations (such as language), in contrast, the schism between behavioral "representation and algorithm" and implementational "virtual machine" arises because language processing involves more than simply mapping inputs to outputs. The system has to store information internally and the nature of the virtual machine doing the internal storing and manipulation has significant consequences for the machine's behavior.

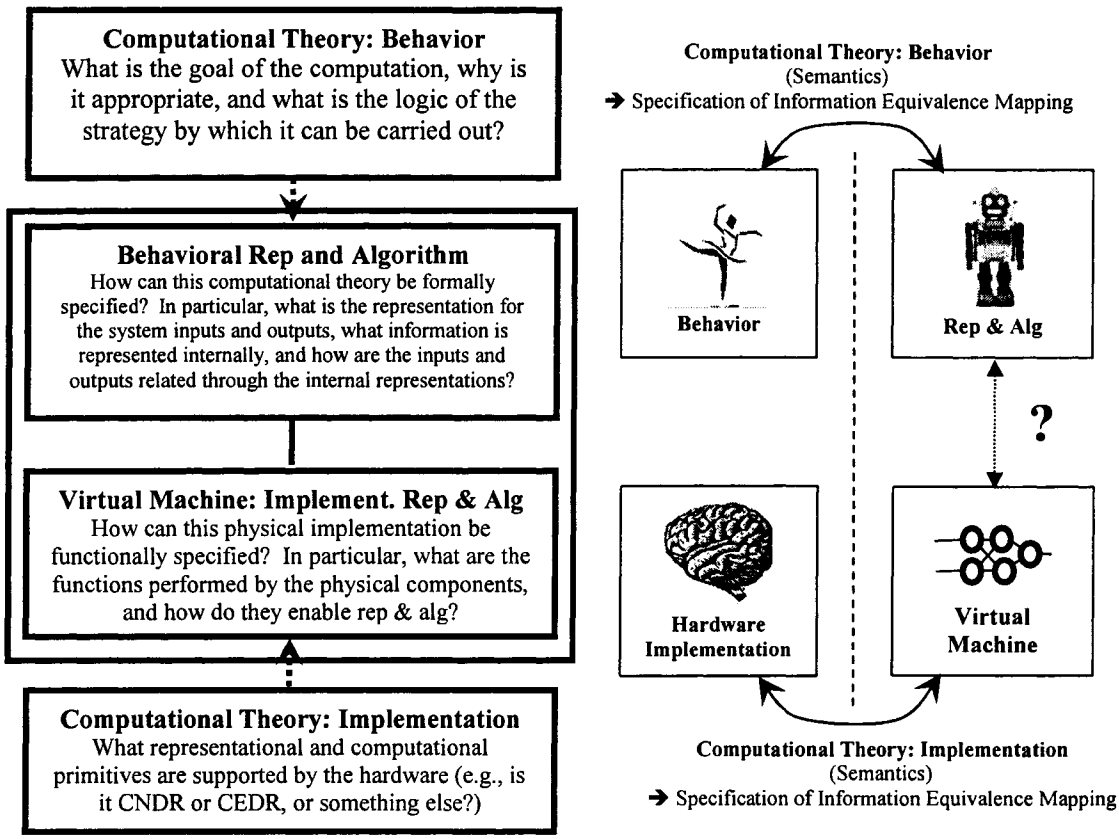
The second problem with this example is that the cash register is a synthetic (or engineered) system that we are free to imagine being designed any way we like, so the hardware implementation description acts like a "free parameter." In particular, the cash register is not a "found" system that we are trying to understand in its own terms. In the case of a found system like the human brain, the implementation level is a constant that we are trying to understand, not a free parameter that we can specify any way we like. In the same way that a behavioral model (i.e., the representation and algorithm description) is specified to capture what is important about the system's real-world behavior in order to understand it better, symmetry suggests that the other major unknown (the components of the hardware) should be mapped onto a functional model (the virtual machine) so it can be understood better, too.

To put it another way, the description of the hardware implementation in Marr's original example (i.e., "Arabic numerals are implemented as 'successive notches on a wheel'") assumes the function of the hardware components are known a priori (in this case, the notches on a wheel are known to represent the Arabic numerals)—indeed, the hardware was designed specifically to implement the "representation and algorithm" specification of addition. In a system like the brain, on the other hand, where the functions of the hardware components (neurons, synapses) are not known *a priori*, a

functional description of the hardware implementation (that is, the virtual machine description) is necessary to pick out the characteristics of the hardware that are assumed to be important, in the same way that the characteristics of behavior assumed to be important are picked out for the representation and algorithm description. In Figure 2.18 I show how the cash register example would be handled in the revised version of Marr's levels of analysis, which include the virtual machine level of explanation.

Note that this analysis suggests an alternate derivation of my analytic framework from a different starting point but leading to essentially the same result shown in Figure 2.3. This derivation starts with Marr's levels of explanation, and treats "behavior" and "hardware implementation" symmetrically since both are referents in the world that need to be modeled and understood, so neither would be included in the levels of explanation framework (just as "behavior" does not appear in Marr's original framework). The "computational theory" is not really a level of explanation at all—it is a specification of the information equivalence mapping between referents in the world and symbols in the model. There would need to be two of these—one for behavior and one for the hardware implementation—to generate the functionalist (behavioral) representation and algorithm description and the physicalist (virtual machine) representation and algorithm description, respectively (Figure 2.19). Marr only included one because the functionalists only deal explicitly with the behavioral side of the system. If the virtual machine were next differentiated into structural and functional levels and the inputs, outputs, representations, and functions at each level were differentiated, the result would be identical to Figure 2.3, but with the addition of explicit terms for the mappings from real world phenomena (behavior, synapses, etc.) to features of the model that represent them. A major question

Figure 2.19: Proposal for a revised version of Marr's levels of explanation based on a symmetrical view of the behavioral and implementational aspects of the neuro-cognitive-behavioral system



in the revised version of Marr's framework becomes how to relate elements of the virtual machine to elements of the behavioral representation and algorithm model. This problem is precisely the one addressed by the scientific method for brain-behavior research I proposed in the introductory chapter of this dissertation.

Discussion

The analysis in this paper is motivated by a desire to conduct rigorous basic research on causal brain-behavior relationships to inform applied educational research, combined with a conviction that computational models can support the crucial step in this endeavor of bridging from brain mechanisms to behavioral patterns. The introduction of computational models into the research process raises a number of challenging questions, however, including:

1) What is the theoretical status of computational models? That is, how should we understand the models as theories of neural and psychological function and observable behavior?
2) How can we identify model properties and behaviors that represent a basis for valid inferences to humans (given model artifacts, model incompleteness, etc.)?
3) How can we verify the models empirically against human data?

I discuss implications of the present analysis for the first two questions in the following sections. The third question is the focus of the next chapter.

Theoretical Status of Computational Models

The first question has no single answer that applies in general because different computational models have different philosophical and theoretical bases, commitments, orientations, and goals. In lieu of a blanket answer, therefore, I have proposed a general

analytic framework that can be applied to any concrete computational model to provide an answer specific to that model. My decision to ground the framework in philosophical materialism was based on two considerations: 1) all major extant psychological and behavioral theories share a belief that mental and behavioral phenomena are ultimately manifestations of neural processes, so this provides a common reference point for comparing and contrasting disparate psychological theories; and 2) the real world is the sphere within which psychological theories, computational models, and the real-world phenomena being modeled all converge, so it is a natural perspective from which to investigate the theoretical status of computational models.

In part to demonstrate the utility of this framework and the process of applying it in specific cases, I analyzed four models of human psychology and behavior: the stimulus-response model, the production system, the perceptron and the multi-layer perceptron. One product of this analysis is a grounded basis for distinguishing between two categories of computational models—functionalist and physicalist—that differ in terms of their theoretical status. Specifically, functionalist models like the production system explicitly eschew theoretical commitments concerning the physical mechanisms that support cognitive processes and behavior. As a result, such models entail mutually exclusive hypotheses about the physical system. Since currently viable functionalist models cannot be falsified based on behavioral tests (they are in principle powerful enough to model any input-output behavior) and they cannot be falsified based on tests of their internal mechanisms (because they make no commitments in that regard), they cannot be evaluated scientifically and therefore should not be considered scientific theories of human cognition or behavior.

In contrast to the functionalist symbolic models, to the extent that physicalist models like the ANN make specific commitments about the physical referents of their model components, they are—in principle—vulnerable to experimental test. In addition, the ANN results make it clear how ANNs bridge from neural mechanisms to behavior, suggesting that these models in particular are good candidates for conducting research linking brain to behavior using the method I proposed in the introductory chapter.

This analysis also generated several by-products. For example, within this framework it became clear why connectionists and symbolists often seem to be talking past one another in the literature—because often they are, in fact, when symbolists adopt a functionist stance regarding symbolic models and connectionists adopt a physicalist stance regarding the same models. In this case, the disconnect occurs because even though the two groups are talking about precisely the same *model* (e.g., a particular production system), they are nonetheless talking about two different *theories* (functionalist and physicalist, respectively) represented by that model. This important distinction between models and theories is supported by the analytic framework introduced here, but not by Marr's "levels of explanation" framework.

I traced the source of this ambiguity to the way Marr's levels of explanation are defined. A second by-product of the present analysis is therefore an identification of some problems with Marr's framework (most importantly, that the "representation and algorithm" level cannot distinguish between qualitatively different implicit theories associated with a single explicit model) and proposals for revisions to help resolve these issues (in this case, by the introduction of an additional "virtual machine" level of explanation).

Finally, the analytic framework is a very general tool supporting direct comparison of disparate neural, psychological, and behavioral theories within a single common frame of reference. By making it possible to distinguish clearly between instances wherein two theories are making common reference to an entity in the world and those where they are not, this tool should be useful for a wide variety of goals beyond those pursued here.

Identifying Candidate Neural Mechanisms

The second question associated with the use of computational models to investigate brain-behavior links is how to identify promising neural mechanisms that can be used as a basis for inference to human neurology and behavior. One strategy demonstrated here is to apply my inter-theoretic framework to compare different theories or models of a common phenomenon and look for conflicts or inconsistencies that might point to mutually exclusive hypotheses that could serve as the basis for an empirical experiment. This strategy is demonstrated, for example, where I identified two mutually exclusive hypotheses about neural architecture as a result of my comparison of the hypothetical physicalist version of a production system and the multi-layer perceptron.

The crucial observation is that distributed representations exist at two levels in the neural system: internal structure and internal activity. In the physicalist production system, these two sets of representations contain the same information—they are basically copies of one another (even though they might have different formats). I call this mechanism "coordinated, equivalent, distributed representations" (CEDR). In the MLP, in contrast, the two sets of representations are coordinated, but they contain different information. I call this mechanism "coordinated, non-equivalent, distributed

representations" (CNDR). These two mechanisms represent mutually exclusive hypotheses about neural organization and behavior. In the next paper I report on an experiment designed to distinguish experimentally between them.

In addition, this concrete case demonstrates the general strategy wherein my analytic framework can be applied to mine the database of extant theories and models in search of candidate neural mechanisms for further investigation.

Conclusions

The analytic framework described here was developed to address specific questions about the theoretical status and research utility of computational models. Primarily owing to its basis in philosophical materialism, however, the framework turns out to be much more widely applicable than I had originally conceived it to be. This generality is evident from the diverse applications and insights derived using the framework even within the context of this paper. These applications include an analysis of individual theories and models; a systematic comparison of disparate theories illuminating a number of interesting similarities and differences and producing a novel taxonomy of psychological theories; identification of problems with Marr's levels of explanation (missing level of analysis, lack of symmetry at the behavioral and implementation ends); and insight into a source of possible disconnect in the current debate between connectionists and symbolists. In addition, I used the framework to identify a potentially important but non-obvious difference between the production system and the ANN which is that the former is not falsifiable while the latter is. Finally, the framework supported my search for hypotheses about neural mechanisms

(CEDR/CNDR) that provide a basis for empirical experiments on causal brain-behavior relationships.

My hope is that through applications such as those described, this framework contributes to the development of a *lingua franca* for comparing and contrasting psychological theories, which will also support the construction of a meaningful taxonomy of psychological theories. One advantage of this particular framework is that it allows for the separation of model from theoretical commitment, so one can infer (or ask the theorist) what the underlying commitments are concerning the material basis of the theory. Once the analysis is completed, two types of theoretical entities can be distinguished: physicalist elements are exposed to falsifiability and functionalist elements are explicitly identified as being non-scientific placeholders until a physicalist commitment can be made. This suggests a more general strategy for making progress in cognitive science: treat functionalist accounts as placeholders until physicalist alternatives can be defined, and then evaluate or differentiate between them empirically when possible. In the next chapter, I attempt to do just that.

Chapter 3

Experimental Paradigm: Quantitative Methods for Testing Causal Brain-Behavior Links

Introduction

Behind any scientific study of education is a deeper question about how people learn. Schools are charged with constantly evolving responsibilities and goals, but ultimately their mission is to facilitate certain learning outcomes in their students. Knowledge acquisition is, at the most fundamental level, mediated by biological processes involving physical changes to a person's nervous system (Bear, Connors, & Paradiso, 1996; Kandel, Schwartz, & Jessell, 2000), so neuroscience seems like a natural place to seek a scientific basis for educational theory and design principles. In recent years, in fact, there has been a great deal of interest in potential applications of neuroscience to education (Bransford, Brown, & Cocking, 2000; Bruer, 1997). Notwithstanding claims by zealous marketers and journalists, however, to date neuroscience has contributed very little of practical value to general education (Bruer, 2002).

A major challenge facing researchers seeking to apply neuroscience to education is the problem of how to establish causal links between neuroscience mechanisms and behavioral patterns, since these phenomena exist at vastly different temporal, spatial, and organizational scales. Computer simulations of brain and mind, such as the connectionist artificial neural network (ANN) model, represent a powerful complement to direct experimentation in this area, because they provide a framework for integrating data from

disparate paradigms and levels of analysis (in particular, from neural mechanism to observable behavior) into a single coherent model, and they allow researchers to simplify and control particular aspects of the model to explore its behavior systematically.

Psychologists have used connectionist models in their research for decades (McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986). The models frequently suggest non-obvious hypotheses about the mechanisms underlying behavioral patterns that can then be translated into the behavioral domain and investigated using standard methods of experimental psychology. However, when evaluating how well a computer model fits empirical data on people's behavior, the dearth of quantitative methods available for formally testing model validity typically forces connectionist researchers to resort to informal comparisons between simulated data and human data[1].

Typically, the logic of ANN model verification is based on positive demonstrations that a particular model is capable of generating a specific empirical data set. Moreover, attempts to falsify the overarching modeling framework itself (in the sense that the stimulus-response framework of behaviorism and the perceptron framework were falsified, for example) using empirical data and quantitative hypothesis tests are very rare, if not nonexistent. Given that educational researchers and practitioners are now making inferences about human learning and knowledge representation from these models and interpreting them to guide pedagogy and

[1] Note that this is not a criticism of either artificial neural network models or the empirical methods employed when investigating hypotheses gleaned from such models. My point here is that the methods whereby model behaviors typically are *related* to human behaviors are indirect and informal, because we lack powerful quantitative methods for *directly* testing model features against empirical human data. Specifically, I am talking about the methods available to ask questions such as: "Is this connectionist network an *adequate* model of that human behavior?" and "Is this ANN a *better* model of the human behavior in question than is a different computational model of the same behavior, such as a production system or a different ANN?" Informal methods (i.e., methods not based on the logic of falsification) are typically used to address such questions.

assessment, we must develop empirical methods for assessing whether such inferences are valid before we gamble precious educational resources on them.

In this paper, I address the broad question, "How can ANN modeling frameworks (like the connectionist framework) be studied using the scientific logic of falsification (in particular, using empirical data and quantitative hypothesis tests)?" To operationalize this question, I have designed an experimental learning task and adapted it for use with connectionist ANNs and human subjects. I used the connectionist simulations to generate two precise predictions about learning behavior on the task, and then I tested those predictions empirically using learning data obtained from a sample of adults performing the same learning task.

Two sections follow this introduction. In the first, I present the theoretical basis of my experiment and describe key aspects of the connectionist model and the predictions I have derived from it that are tested against human learning data. Following this, I describe the critical features of the experiment, including my sampling, procedures, measures, data analyses, and findings.

Background and Motivation: Computer Simulations and Learning Research

A major challenge facing researchers seeking to apply neuroscience to education is determining how to establish and validate causal links between neural mechanisms and behavioral patterns, since these phenomena exist at vastly different temporal, spatial, and organizational scales. This problem is compounded by the fact that invasive controlled experiments, which represent the most direct route to investigating brain-behavior relationships, cannot be conducted on people. While such experiments can be conducted using animal models, the higher cognitive functions most obviously relevant to education

(such as language and mathematics) are unique to human beings, and as a result animal research can offer little—if any—insight into the neural mechanisms supporting these complex functions and their associated behaviors.

Direct experimentation is complemented powerfully by computer simulations in many research domains, such as astronomy, meteorology, economics, physics, mathematics, chemistry, and biology. Computer models become doubly important in domains where the phenomena are complex and direct experimentation is difficult or impossible, as in brain-behavior research in people. One computational model derived from neuroscience commonly used in psychology is the connectionist artificial neural network (ANN). In this modeling framework, researchers use networks of processing elements (nodes) inspired by biological nerve cells (neurons) to construct learning systems that can master a wide variety of tasks (cf. McClelland & Rumelhart, 1986).

In recent years, educators have begun using connectionist models to reason about human cognition (Bereiter, 1991; Schneider & Graham, 1992), making recommendations about pedagogy (Baker & Martin, 1998; Jones, Hill, & Coffee, 1998; Roth, 1992) and assessment (McKnight & Walberg, 1998; Papa, Shores, & Meyer, 1990; Perkins, Gupta, & Tammana, 1995) based on network behavior. This shift from the theoretical domain of psychology to the more applied domain of education raises important issues concerning the validity of the connectionist ANN as a model of learning processes and knowledge organization in people. These issues have always lurked in the background of ANN research, but they quickly come to the fore when educational researchers propose to use these models to inform the design of educational materials, strategies, and environments

to be used in schools, where the stakes are tangible and substantial. In my view, two issues in particular become paramount in this context.

The first problem pervasive in the literature on ANNs and education stems from the tendency to take the computational model as a starting point and freely interpret any and all properties and behaviors of the model as representative of neural processing in people (for example, see Andersen, 1999; Anderson & Donaldson, 1995; Anderson, 1992; Anderson & Conway, 1997; Baker & Martin, 1998; Baker, 1994; Roth, 1992). The following example is typical of this process.

Roth (1992) analyzes the behavior of several connectionist ANNs learning to solve balance beam tasks. Based on his analysis, he identifies a number of model properties that appear to correlate with findings in science learning research:

> These [balance beam] simulations show the following properties of ANN models that are consistent with recent work on science learning but not yet sufficiently appreciated by many science educators and teachers.
> 1. Consistent with recent work on the relationship between rules and situated action..., ANNs learn to correctly solve problems without explicit rules as causal determinants of responses.
> 2. Consistent with work on scientists' perceptions..., ANNs learned to perceive in prototypical ways over time and in incremental ways.
> 3. Consistent with work in cognitive science, artificial intelligence, and bottom-up robotics..., ANNs enact knowledge rather than store it. Patterns that correspond to knowledge are not found in an ANN, but in the activations that propagate through it.
> 4. Consistent with psychological experiments involving human subjects..., ANNs show incremental and proximal development.
> 5. Consistent with my own observations of students in science laboratories..., ANNs abstracted different patterns—that is, learned different concepts from the same set of materials (p. 72).

Roth then uses these surface correlations between ANN model properties and observed human behavior to make recommendations for improving science teaching.

I do not mean to suggest that there is anything wrong with this particular example, or even with the general strategy it exemplifies. Indeed, in my opinion Roth's analysis is

one of the more thoughtful and interesting applications of ANNs to educational practice that I have seen. My point is that the analytic strategy exemplified here is very weak because it is based on surface correlations without any kind of formal validity test to back up the interpretations being made. At present, this strategy is very common because formal alternatives do not exist.

The problem with the informal strategy exemplified above is that while some model properties are likely valid representations of corresponding neurobiological properties, other model properties are definitely artifacts of the way the model is constructed and bear no relation to the neurobiological system. There is no principled way to distinguish the valid from the artifactual properties from within the model itself. That is, when researchers take a specific ANN model (or the more general modeling framework) as their point of departure for drawing inferences to people (as in the example), they have no grounded basis for distinguishing valid model properties from invalid model artifacts. It becomes very difficult (if not impossible) to make careful inferences from model behavior to human behavior under these conditions.

I wish to be very clear here. I am not saying that inferences from ANN models to human neurobiology using such methods as I have just described are necessarily false. Nor am I arguing that the conclusions and interpretations derived from such inferences are necessarily wrong. My point is that these methods provide no direct evidence one way or the other supporting reasoned judgments about the truth or falsity of the conclusions. That is, the surface correlations and the conclusions drawn from them are just as likely to be false as they are to be true (if not more so). Without some kind of

validity check, these "conclusions" are really "hypotheses" that require substantiation and validation before they should be used to inform educational designs.

On the positive side, it is possible that these kinds of informal and exploratory methods can produce novel insights and associated hypotheses that would not have been discovered using behavioral data alone (for example, by suggesting the kind of mechanism implicated in specific behavioral patterns such as those enumerated in the example above). In addition, such methods can be used as "proofs-by-example" to demonstrate that a system like an ANN is sufficiently powerful to produce a particular pattern of behavior that might seem to require more specialized—even innate—structures and processes (such as embedded clauses in language—see, for example, Elman, 1995). If they support novel or improved educational designs, then the models could provide substantial practical benefits in education even when such informal methods are used. Without more rigorous arguments linking specific model properties to neurobiological mechanisms and more powerful methods of validation, however, model behaviors and properties are merely suggestive and speculative—they cannot really be taken seriously as *evidence* for or against any particular explanatory theory of neural function or behavior.

The second consideration from the standpoint of scientific validity is that connectionist researchers typically resort to informal (that is, *ad hoc* or *post hoc*) comparisons when validating their computer models against people's behavior (see, for example, Elman et al., 1996; McClelland & Rumelhart, 1986; McLeod, Plunkett, & Rolls, 1998; Quinn & Johnson, 1997; Rumelhart & McClelland, 1986), rarely designing true experiments, collecting quantitative data and conducting formal tests in an effort to

falsify the hypothesized model—or, even more powerfully, to falsify the overarching modeling framework. In fact, there do not seem to be any general methods for testing the validity of ANN models in this way. This issue is linked to the previous one, in that before one can conduct a meaningful hypothesis test, one must first be able to state a meaningful, plausible, and testable hypothesis.

In this paper, I propose novel experimental methods for testing whether the simulated learning processes and internal representations of ANNs are indeed similar to those used by people. The general experimental approach involves four steps: 1) identify a neural mechanism, 2) embed the neural mechanism in an artificial neural network model, 3) generate behavioral predictions from the computational model, and 4) test the model predictions using data from human subjects. In this context, the computer simulations permit me to make precise *a priori* predictions. If such predictions are supported empirically, in an experimental setting, then the ANN model and/or modeling framework become more compelling and we gain insight into human cognition. If the predictions are not supported, then we learn something about the strengths and limitations of ANN models and can refine them. Either way, experiments such as the one I report on could inform our use of ANNs to generate cognitive theory and inform educational practice.

Identifying a Neural Mechanism: Coordinated, Non-Equivalent, Distributed Representations (CNDR)

ANN researchers make much of the fact that ANNs employ (spatially) "distributed representations," partly because this also seems to be a property of representations in the brain (Elman et al., 1996; Klahr & MacWhinney, 1998; McLeod et al., 1998). For example, consider an ANN that learns to take the present tense form of a

verb as its input and produce the appropriate past tense form as its output. In the network, the knowledge necessary to convert a present tense form like "ring" into its past tense form "rang" would be distributed throughout the synaptic weights and node thresholds of the network instead of being stored in any one place. A "localist" representation of the rule for converting "ring" to "rang," in contrast, would be stored in a single location separate from all other information (as it is, for example, in a grammar textbook). The problem is, it is difficult to define precisely what differentiates distributed representations from localist representations as defined in this way. Localist representations (such as the words on the page of a grammar textbook) are also spatially distributed, so being distributed in itself is not a distinguishing factor. Furthermore, many people doubt the biological plausibility of distributed representations as represented in connectionist ANNs (Elman et al., 1996; Klahr & MacWhinney, 1998). For these reasons, attempts to differentiate ANNs from other kinds of computational models (such as production systems) based on their distributed representations has not been very successful.

I argue that it is not the presence of distributed representations *per se* that is most interesting or important about ANNs. ANNs, like biological networks, employ two distinct types of distributed representations—distributed *weights* and distributed *activations*. When researchers refer to distributed representations in ANNs, they rarely specify which type they mean, although it can be inferred from context that they usually mean distributed patterns of activation (the basis for applied knowledge). In my view,

what is interesting (and biologically plausible) about ANNs is the way these two types of distributed representations are *coordinated*[2].

In particular, there are two possible ways to coordinate two sets of distributed representations[3]: 1) make the two sets of representations contain the same information (i.e., make them copies of one another), or 2) make the two sets of representations contain different information. To illustrate these two possibilities, imagine you have collected data on shoe size and math achievement from a number of children (Table 3.1).

Table 3.1: Imaginary data on children's shoe sizes and performance on a math achievement test

Subject ID	Shoe size (in.)	Math achievement
1	4 in.	8%
2	5 in.	19.5%
3	7.5 in.	48.25%
4	8.25 in.	56.875%
5	12 in.	100%

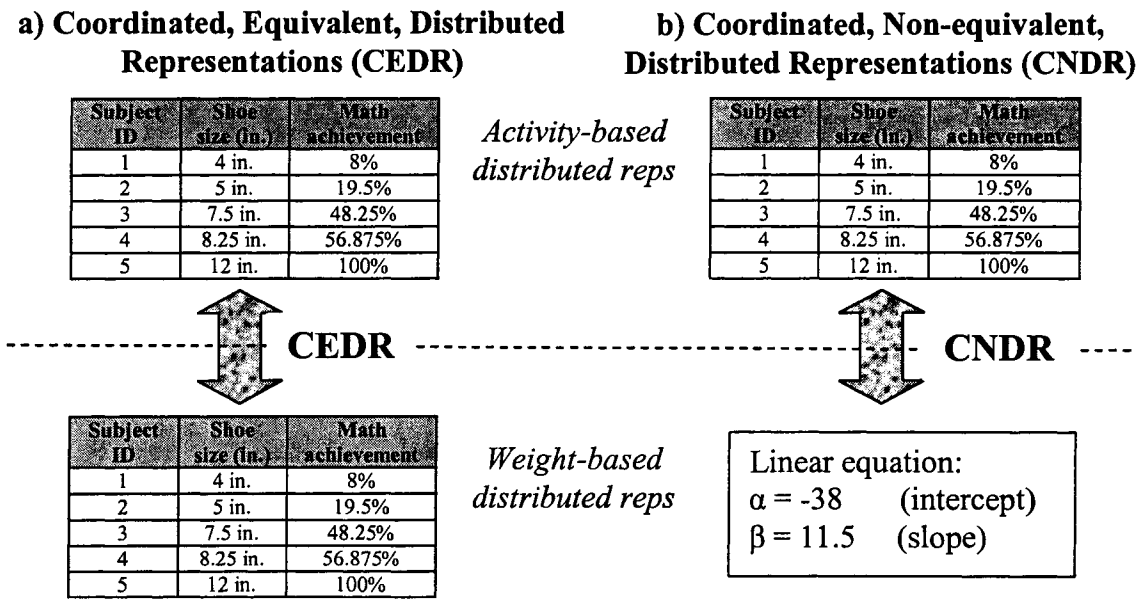
Using two sets of distributed representations, there are two ways these data could be stored:

Coordinated, Equivalent, Distributed Representations (CEDR): In this case, the two sets of distributed representations contain the same information, although perhaps in different formats (see Figure 3.1a). For example, if the data in Table 3.1 were entered into a spreadsheet on a computer, the spreadsheet loaded into RAM (working memory) would be like the activity-based representation. If the file were then saved, an exact copy of the spreadsheet contents would be copied from working memory onto the hard disk (long-term memory). The file stored on the computer's hard disk would be analogous to the weight-based representation. In this scenario, these two sets of representations

[2] See Bishop (1995) for a technical discussion of these different types of representations, which he calls *non-parametric* and *semi-parametric*.

[3] See chapter 2 for a more thorough derivation of this argument and the CEDR and CNDR mechanisms, and see chapter 4 for a more informal and accessible treatment of them.

Figure 3.1: Two ways to coordinate two sets of distributed representations: a) both sets contain the same information (CEDR), or b) the two sets contain different information (CNDR).



contain identical information—both store the set of data points listed in Table 3.1. These two sets of representations (the spreadsheet in RAM and the spreadsheet file on hard disk) are, for all intents and purposes, copies of one another[4].

Coordinated, Non-equivalent, Distributed Representations (CNDR): In this scenario, the two sets of distributed representations contain different information (which goes beyond a superficial difference in formats). For example, the data in Table 3.1 exhibit a perfectly linear relationship, so the contents of the table can be summarized without error using the linear equation:

$$Math\_achievement = -38 + 11.5*Shoe\_size$$

This equation implicitly contains all of the information in Table 3.1 (and in addition, note that it automatically generalizes to other values of shoe size not in the original data set). When we want to work with the shoe size and math achievement data (in the activity-based representations) we need access to the numbers themselves, but this does not mean that we have to store those numbers in the weight-based representations for direct recall. Instead of storing the table itself we could just as easily (if not more easily) store the linear equation parameters (intercept=-38 and slope=11.5) in the weight-based representations and use these to generate values of Math_achievement on demand. These two sets of information (the slope and intercept on the one hand and the set of shoe size and math score pairs on the other) are clearly coordinated with one another, but they are also obviously not copies of one another (note that none of the numbers in the table bear

[4] Note that connectionists would classify these two representations as "localist" instead of "distributed" because the document contents are stored more or less in one contiguous region of RAM and/or hard disk, and each document is stored separately from all others. For present purposes, I want to shift attention from the "distributed" vs. "localist" debate (which I consider a red herring) and therefore I am defining "distributed" in the present context to mean simply "spatially distributed."

any relationship to the values of the slope and intercept)—they contain different information (Figure 3.1b).

I submit that the CNDR mechanism is a property of biological neural networks that is already embodied faithfully in the connectionist ANN model. In order to ameliorate the difficulties involved in inferring from model behavior to human behavior mentioned above, I propose to identify model properties and behaviors that are consequences of this mechanism and base my predictions about human learning only on them.

Generating Behavioral Predictions from the CNDR Mechanism Embedded in the ANN Model

Similarity Structure and the Shape of Learning in ANNs

ANN learning behavior is influenced by relationships among the items being learned, what researchers call the "similarity structure" of the task (see Elman et al., 1996; Plunkett & Elman, 1997). Similarity structure has been implicated in numerous cognitive processes (Goldstone, 1999; James, 1890), including analogical reasoning (Gentner, 1983), concept formation and categorization (Goldstone, 1994), and knowledge transfer (Fischer & Farrar, 1987; Salomon & Perkins, 1989). As a result, many psychological theories incorporate some notion of similarity.

Many similarity-based theories have been criticized, however, for failing to specify an objective *a priori* similarity metric independent of subjective human (typically *a posteriori*) similarity judgments (Goldstone, 1999; Goodman, 1972; Salomon & Perkins, 1989). For example, in knowledge transfer research, the "distance" of transfer is typically determined by the researcher's subjective evaluation of how similar two application contexts are. Transferring knowledge of driving from a car to a rental truck is

considered *near* transfer because the contexts are similar, whereas transferring strategies from chess to the business domain is considered *far* transfer (Salomon & Perkins, 1989). Although these distinctions make intuitive sense, they can be problematic in practice because similarity in the outcome depends upon measures of transfer distance that are themselves based on subjective similarity judgments, creating a vicious circle.

Connectionist models provide new tools for posing questions involving similarity, allowing researchers to define the similarity structure of a stimulus set objectively, without reference to subjective human judgments, and then to ask questions about the relationship between this objective baseline and the subjectively perceived similarity structure. The relationship that a connectionist network establishes between objective and subjective similarity structure on a task is crucial to its performance and profoundly shapes its characteristic learning behavior. Moreover, the form this relationship takes is a direct consequence of the CNDR mechanism. I therefore exploit this feature of connectionist models (which appears to correlate with a property of biological neural networks) in my experimental design by identifying specific behavioral predictions that derive from it.

I use the connectionist simulations as the theoretical basis for generating two specific hypotheses about human behavior, one relating to the role of objective similarity structure in determining task complexity, and the other pertaining to the way subjective perceptions of stimulus similarity change as learning proceeds. Two challenges arise in connection with this goal. First, I must establish a framework for relating simulated data to empirical data from people. Second, I must define the similarity metric I propose to use to formalize my experimental hypotheses.

Relating Simulation Behavior to Human Behavior

Connectionist model construction involves designing a stimulus set for use in training the ANN to perform a target task (Plunkett & Elman, 1997). For example, McClelland (1989) constructed a network that learned to solve balance beam problems (Inhelder & Piaget, 1958), and Seidenberg and McClelland (1989) designed a network that learned to "read aloud." Basically, to achieve this, the modeler establishes a set of rules for converting human-friendly stimuli (balance beam problems, letter strings) into computer-friendly numerical codes.

As a concrete example of this process, consider the cartoon face in Figure 3.2. This face can be characterized by the horizontal distance between the eyes (E) and the vertical extent of the face from lips to eyes (F). Beginning with this prototype, I systematically varied the values of E and F to create a set of related but distinct faces—a "face-space" (see Figure 3.3 for a caricature of the face space; the actual stimuli used in the experiment can be found in Appendix A). The values of E and F for each face serve as the numerical representation of the stimulus used to train the neural network.

In this example, I have specified a rule for converting a pictorial stimulus (a particular face) into the corresponding network input (numbers uniquely identifying that face) and vice-versa. This is a standard approach in connectionist modeling (Elman, 1990, 1993, 1995; Elman et al., 1996; McClelland & Rumelhart, 1986; McLeod et al., 1998; Oliver, Johnson, Karmiloff-Smith, & Pennington, 2000; Plunkett & Elman, 1997; Quinn & Johnson, 1997). I used the actual face stimuli with human subjects in a learning task and the corresponding set of numerical stimuli with the connectionist network in an analogous simulated learning task. The rule that relates one to the other enables me to

Figure 3.2: Prototypical experimental stimulus

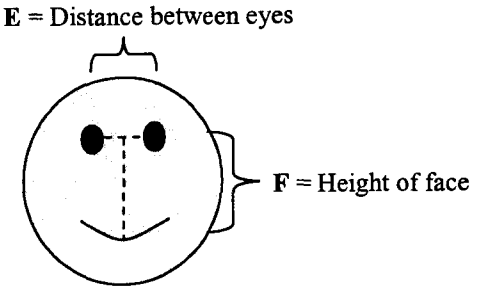
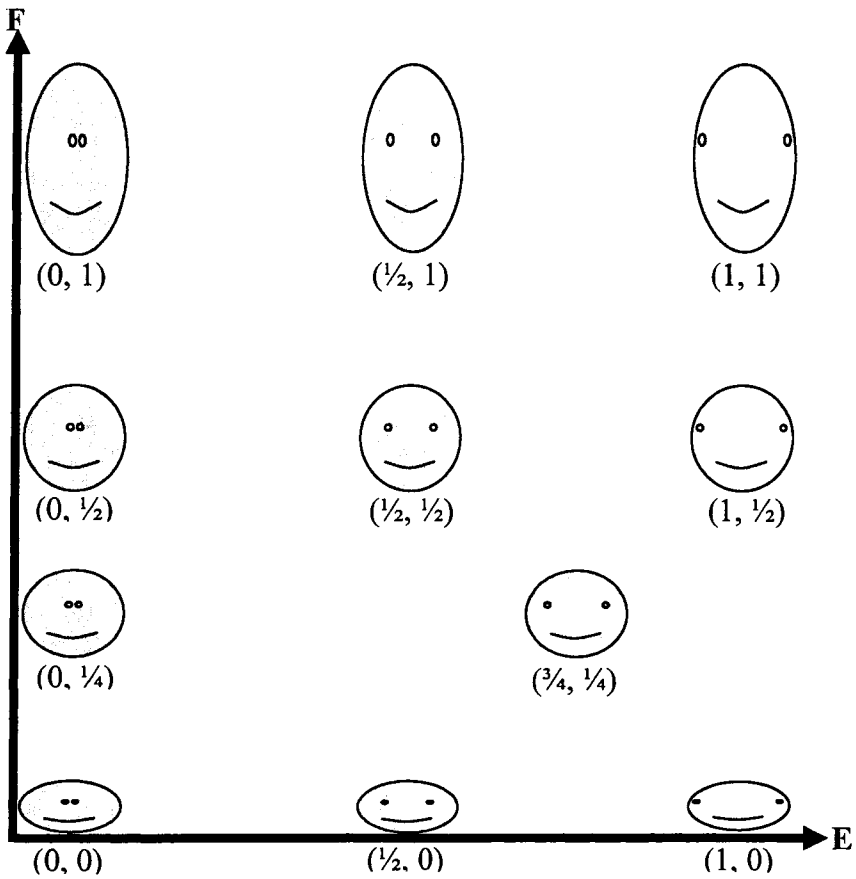


Figure 3.3: The "face-space" relating human stimuli to ANN stimuli



translate between the simulated and real worlds and thus compare simulated to human data.

Defining an Objective Similarity (or Difference) Metric on the Face Space

Once the critical variables characterizing a stimulus have been identified and mapped to numerical values, it is straightforward to define a similarity metric on the set of stimuli. Since each stimulus can be identified by a point (E, F) in face-space, researchers often define the "difference" between two stimuli (which is the inverse of their "similarity") as the Euclidean "distance" between the two corresponding points (Figure 3.4). In designing my experiment, I used this similarity construct to formalize and test two model predictions about learning and knowledge representation as described in the following sections.

Overview of the Model Predictions

To generate precise predictions that can be tested experimentally, I trained a connectionist network to perform a simple task. First, I divided the face-space arbitrarily into two halves to represent two types of imaginary creatures, called "Gorfs" and "Dimps" (Figure 3.5). The object of the task is to learn through experience with feedback to correctly identify the group membership of each stimulus (similar to the way a biology student might learn to distinguish frogs from toads by studying many labeled examples of each). On each trial, a face is presented to the network, the network indicates the type of creature it thinks the face represents, the correct answer is presented as feedback, and the network accommodates the feedback through incremental learning. Eventually, the network exhibits mastery of the task by categorizing all the faces correctly. In my analyses, I looked beyond this end-state performance to investigate

Figure 3.4: Computing a Euclidean distance in face-space. The smaller the difference between two faces, the greater is their similarity.

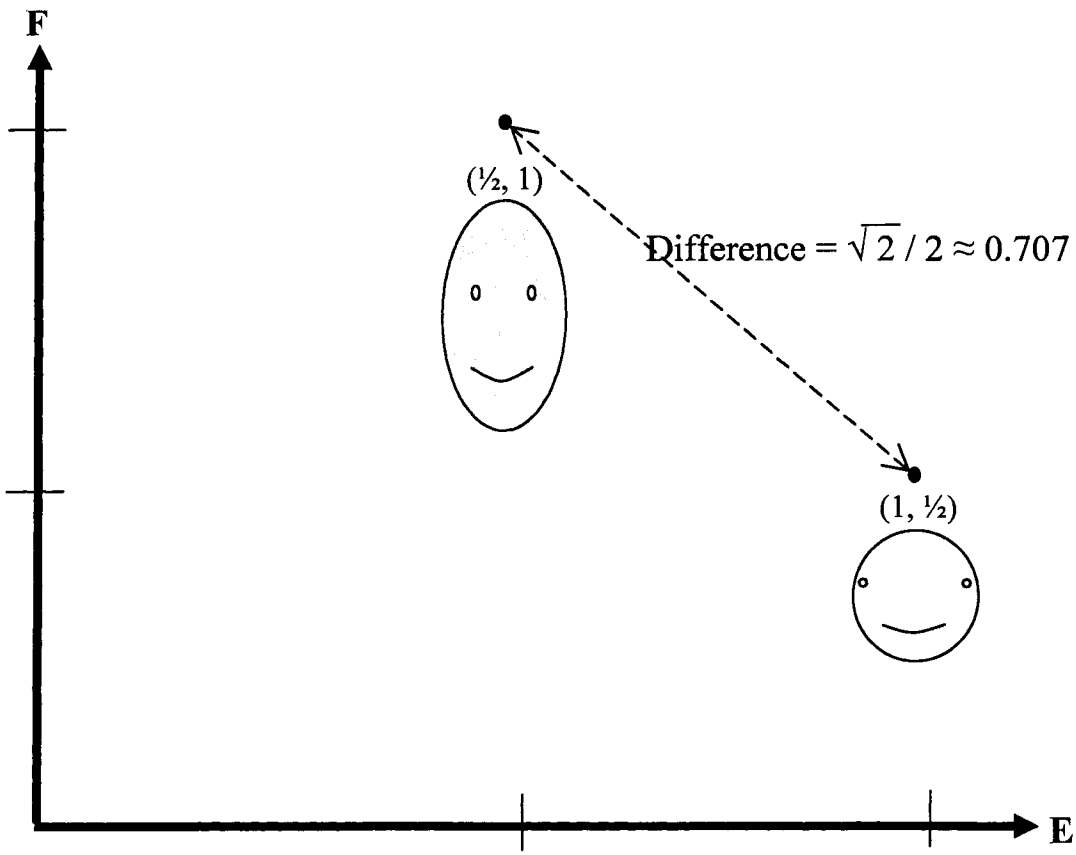
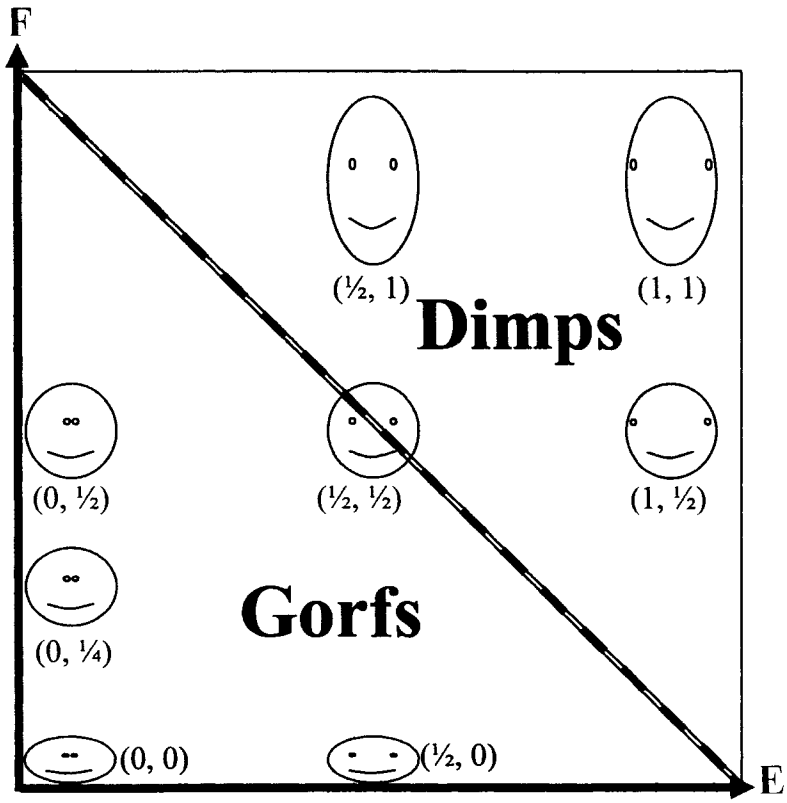


Figure 3.5: Face-space is divided to define two species of imaginary creatures



whether the simulation follows a learning trajectory and constructs internal representations similar to those of people. To investigate these questions, I examined the behavior of the simulation and from that examination I generated two non-intuitive behavioral predictions causally related to the CNDR mechanism. I then tested these predications empirically in human learning.

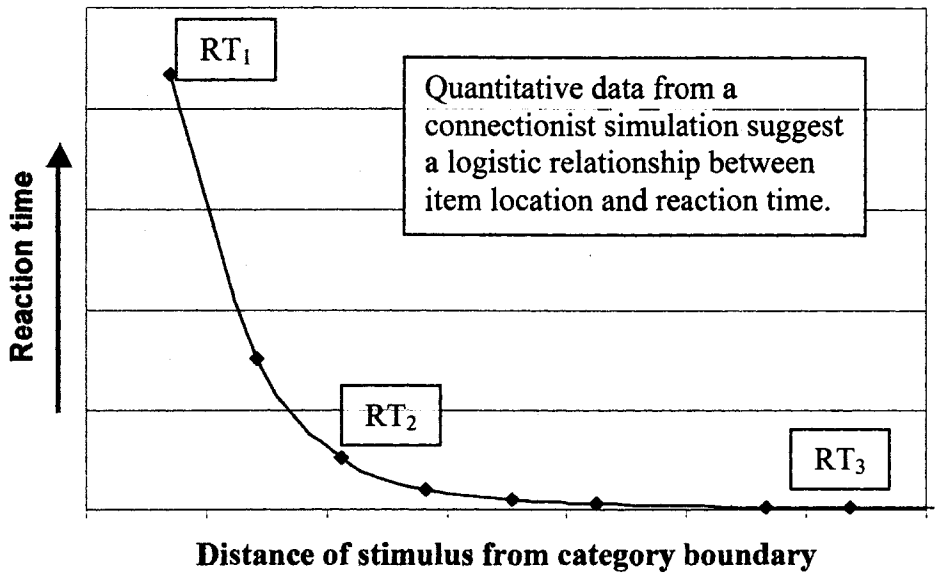
Simulation Prediction #1: Some Stimuli Are Harder to Learn than Others

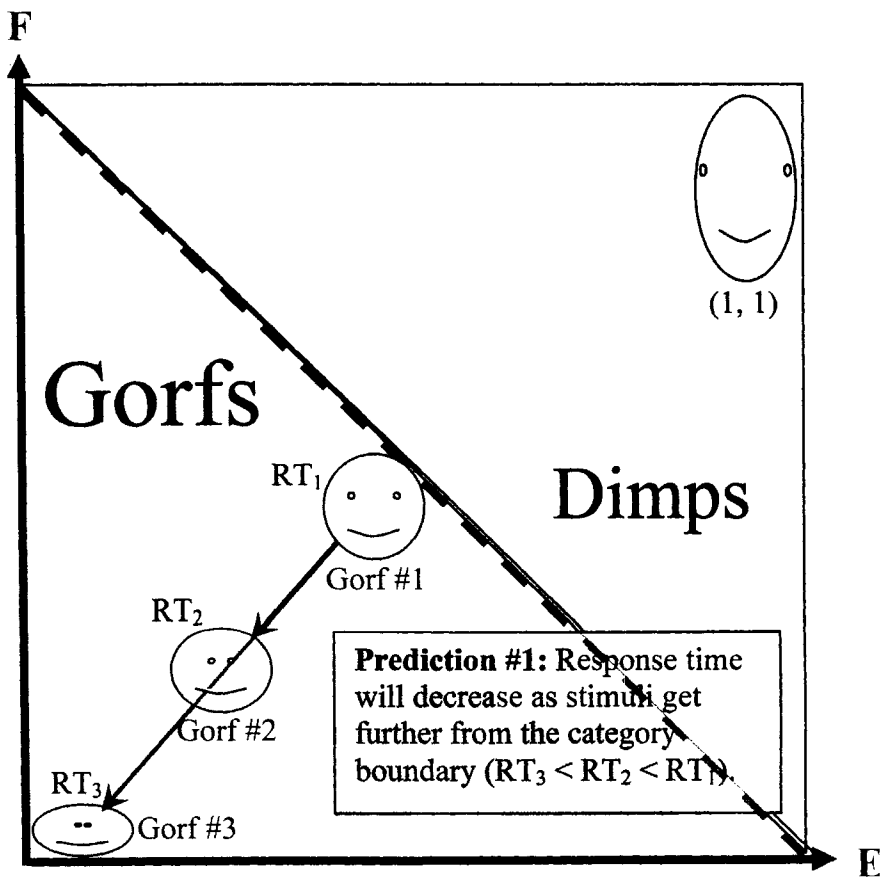
Why are some things harder to learn than others? The answer is undoubtedly very complex, involving many interacting factors. Cognitive simulations can help us identify the individual factors involved and isolate them for focused study. For example, we might ask whether some faces in Figure 3.5 are harder to learn to categorize than others, and if so then what determines their relative complexity. At the outset, there are many plausible hypotheses, including one that predicts all the faces will be of equal difficulty. Based on my analysis of the connectionist simulation, I predict that stimuli near the category boundary are harder to learn than stimuli further away (see Figure 3.6), where distance is measured using the difference metric described previously, and subject reaction time is taken as an index of item difficulty (as is common—see Posner 1989).

The simulation makes a precise prediction about the relationship between reaction time[5] and stimulus distance from the category boundary (Figure 3.6, top panel), during learning and even after mastery. In the simulation, reaction time drops off as the lower half of a negative logistic ("squashed-S") function with distance from the category boundary. This specific functional form, however, depends on "mechanical" details of

[5] The quantity graphed on the vertical axis in Figure 3.6 is not actually reaction time. It is a quantity called the network "error," which reflects the certainty associated with the network's response to a particular stimulus. Connectionist researchers generally assume this error measure is related to reaction time in people (McLeod et al., 1998; Seidenberg & McClelland, 1989), and I make the same assumption here.

Figure 3.6: Predicted relationship between stimulus location in face-space and reaction time. The bottom panel is a graphic interpretation of the quantitative network data plotted in the top panel.



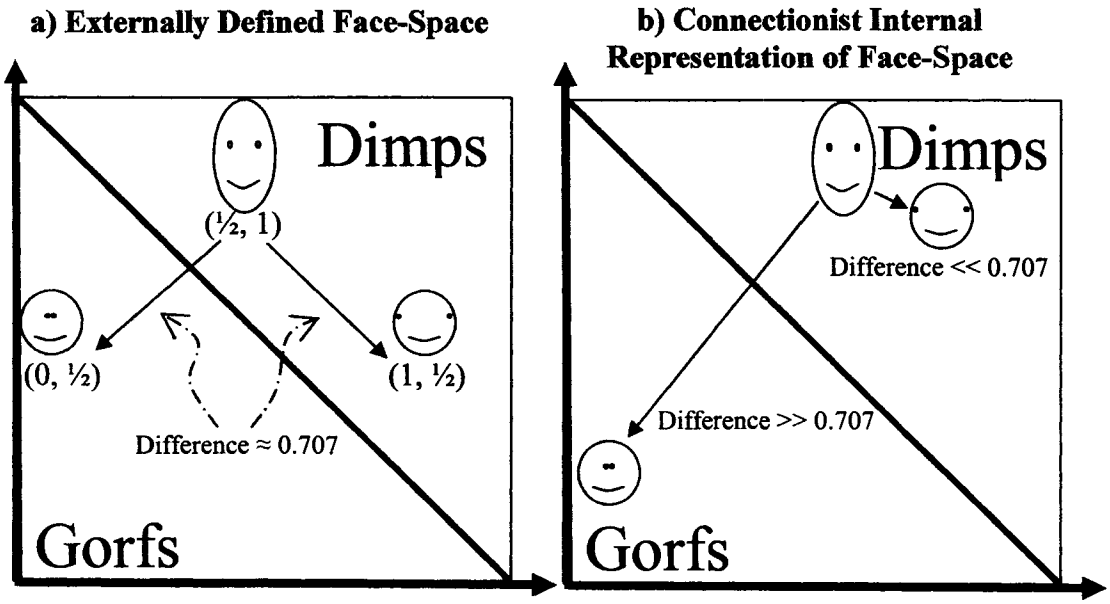


the specific connectionist model that may not be mirrored identically in people (Anderson, 1995; Rolls & Treves, 1998). In addition, the state of the model can be "frozen" while its internal representations are probed to reveal this precise shape, whereas the internal state of human subjects is assumed to be in a constant state of dynamic activity and subject to various sources of noise not present in the model. Initially, in my experiment, therefore, I first tested the hypothesis that on average, reaction time decreases with distance of the stimulus from the category boundary using a hypothesized linear model as a first approximation to the reaction time/distance relationship. I then examined the functional form of the relationship more carefully in an effort to infer something about its shape.

Simulation Prediction #2: Perceived Similarity Changes Systematically with Learning

While the first prediction addresses factors contributing to item difficulty, the second prediction pertains to the cognitive mechanisms actually involved in knowledge construction. That is, when people acquire new knowledge, how is that knowledge organized internally by the cognitive system? In the face-identification task, for example, each face and label might be stored as associated pairs of isolated facts, much like names and addresses in a personal organizer. The connectionist model, however, suggests a very different kind of organization, in which the critical structures are not the individual facts (faces and labels), but the face-space as a whole (Elman et al., 1996; Plunkett & Elman, 1997). The network exploits the fact that all the Gorfs are in one region and Dimps in another by constructing an internal version of face-space that is "warped" along the category boundary, pushing Gorfs one way and Dimps the other (Figure 3.7). This behavioral pattern is another direct consequence of the CNDR mechanism.

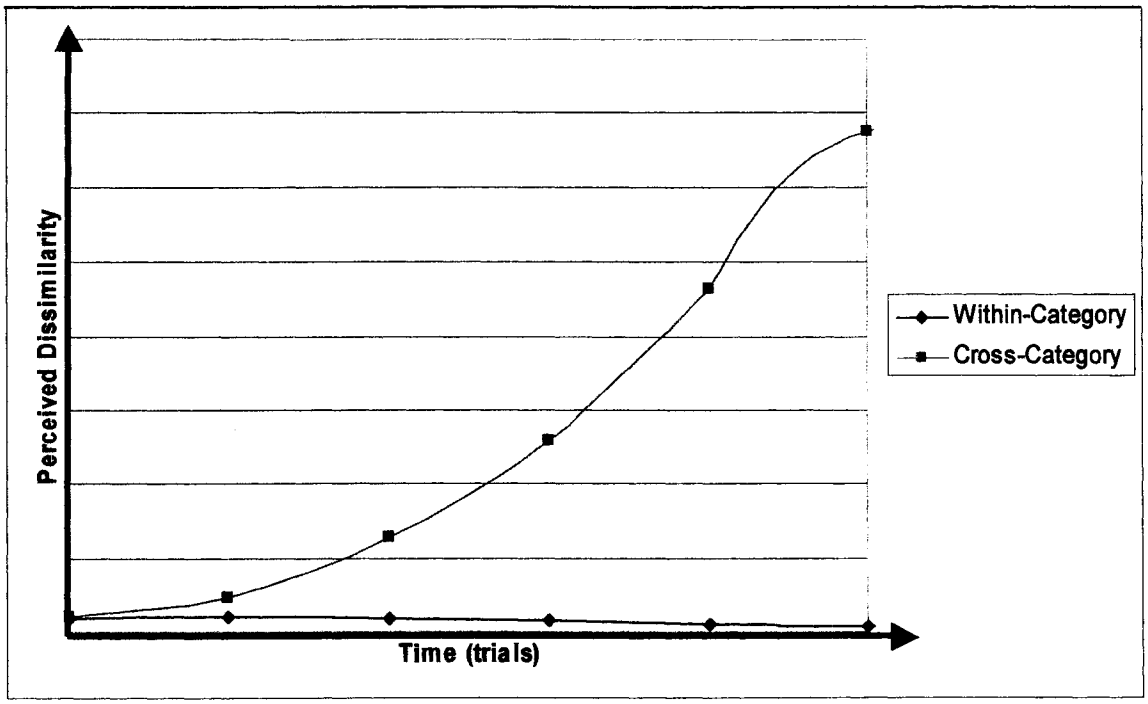
Figure 3.7: Pairs of faces that are equally similar in face-space (left panel) are not represented internally as being equally similar by the connectionist simulation (right panel). The simulation masters the task by constructing an internal representation of face-space in which within-category differences are reduced and cross-category differences are exaggerated.



Assuming that the biological equivalent of the network's internal representations are the basis for our perceptions, the simulation predicts that stimuli within a single category will come to be perceived as increasingly similar to one another and increasingly different from stimuli in the alternate category as learning proceeds. To demonstrate this, I selected a target face from one category and two equidistant comparison faces—one from the same category and one from the alternate category (see Figure 3.7a). At intervals during the training of the neural network, I probed the simulation's internal representations to determine its perceptions of the relative similarities of these two stimulus pairs. I found that the within-category difference becomes smaller and the cross-category difference becomes greater as learning proceeds (Figure 3.7b illustrates this abstractly, and Figure 3.8 plots actual simulation data).

Researchers have suggested that these simulated internal representations are like our own (Elman et al., 1996; McLeod et al., 1998; Spitzer, 1999). In my experiment, therefore, I investigated whether people's perceptions of stimulus similarity change during learning in this way. The upper learning trajectory in Figure 3.8 looks a bit like a squashed letter "S" (i.e., it appears to have an approximately logistic functional form). In the previous case (Figure 3.6), the shape of the curve depended on the shape of the simulated node's activation function. In this case, in contrast, the hypothesized functional form seems to depend more on the learning process than on specific structural details of the simulation such as the node activation function. However, the maximum duration of a single data collection session with humans is limited by exhaustion and boredom, so in this experiment only a part of the trajectory is traversed, which is likely to be approximately linear if this prediction is correct. In the analysis, therefore, I first test

Figure 3.8: The connectionist simulation predicts that as learning proceeds, stimuli within a category will be perceived as increasingly similar to one another and increasingly dissimilar to stimuli in the alternate category



using a hypothesized linear model, and then I look more closely at whether learning proceeds with a logistic (squashed "S") trajectory.

Conclusion: Specific Research Questions

Broadly speaking, I would like to evaluate the validity of ANNs as tools for modeling human cognition using empirical data, quantitative methods, and the logic of falsification. In this initial study, I have used the connectionist framework to generate two testable predictions about human learning and knowledge. These predictions lead directly to two specific research questions that I have addressed empirically in a sample of human adults:

- RQ1: After some learning has occurred on the categorization task in human adults, do reaction times decrease systematically with stimulus distance from the category boundary?
- RQ2: As human learning proceeds, do stimuli within a category come to be judged as increasingly similar to one another and less similar to stimuli in the alternate category?

Of course, there are important differences between the simulation and living subjects that could affect my results in ways unrelated to my research questions. First, I administer the experimental tasks via computer, which could add a level of complexity compared to a non-computerized version of the task for people less familiar with computer interfaces (Chua, Chen, & Wong, 1999; Goldberg, 2000; Rozell & W. L. Gardner, 1999). Second, research has revealed gender-related differences in computer use (Chua et al., 1999; Li, 2002; Turkle & Papert, 1991). Third, research has shown age-related effects on memory (Craik, 1986; Mead, Batsakes, Fisk, & Mykityshyn, 1999;

Mitchell, Brown, & Murphy, 1990) and computer interaction (Chua et al., 1999; Czaja, 1996; Mead et al., 1999) that could influence my results. I therefore controlled for these effects in my data analyses.

Research Design

Sample

I recruited a convenience sample of 48 adults (26 female, 22 male) for the experiment (the recruiting flyer is in Appendix B and the permission form is in Appendix C). A convenience sample is appropriate since I am investigating what many researchers believe are universal properties of human cognition (Elman et al., 1996; McLeod et al., 1998; Spitzer, 1999). For the same reason, I did not anticipate systematic effects of major background variables. Nonetheless, for reasons discussed in my literature review, I made an effort to recruit a sample diverse with respect to gender, age, and computer experience to ensure that any such variation is represented in the sample (the background questionnaire used to elicit this information is included in Appendix D).

In order to avoid excessive between-subject differences in learning due to ongoing cognitive development, I drew a sample of adults over the age of twenty-one, as developmental psychologists have established that by this age people are generally capable of formal or abstract reasoning (Fischer & Bidell, 1998; Flavell, Miller, & Miller, 1993; Gruber & Voneche, 1995) and have undergone substantial frontal lobe maturation associated with executive function (Crown, 1996; Stuss, 1992).

Power analyses conducted with the OpDes software (designed by Congdon and Raudenbush, 2001, for conducting power analysis in multilevel analyses) suggested that a sample size of 50 subjects would permit me to detect moderate effects with a statistical

power of .80 at the .05 alpha-level (Raudenbush & Liu, 2001). Both experimental tasks were implemented using the DMDX software package (Forster & Forster, 2003).

Prediction #1: Item Difficulty Varies as a Nonlinear Function of Distance from the Category Boundary

Procedures

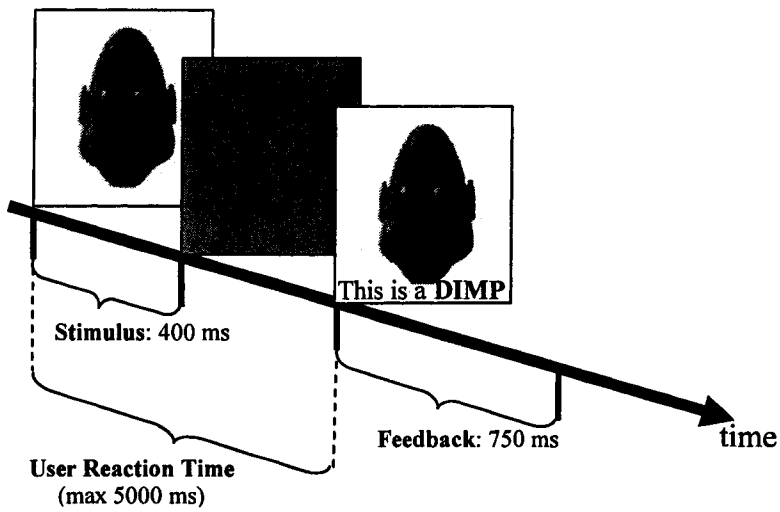
I devised a game-like task (described previously) in which a subject must learn category membership ("Gorf" or "Dimp") for each face in a stimulus set through trial-and-error. On each "categorical" trial (Figure 3.9), a randomly selected stimulus is presented via computer to a subject briefly (400 ms). The subject responds by pressing one of two buttons to indicate their response. If the subject does not respond within 5 seconds, the trial times out and is logged as "no response." After each response (or timeout), the subject receives feedback indicating the correct response (750 ms). This procedure is repeated for four hundred trials (divided into four equal blocks, with five presentations of the entire set of twenty stimuli in each block).

Measures

Outcomes

In addressing RQ#1, I am testing whether some faces are harder to categorize than others. On each categorical trial, therefore, I recorded the subject's reaction time (RT) in milliseconds as the outcome measure. The simulation predicts a nonlinear reaction time curve with characteristics of a decreasing exponential (see Figure 3.6), and therefore I transformed the reaction time data using the natural log function to linearize this variable. In all of my analyses for this research question, I used ln(RT) as the outcome variable.

Figure 3.9: Structure of a category learning trial. A face is presented for 400 msec and then blanked out. Subject reaction time is measured from the time of stimulus onset until a key is pressed. After user responds, the correct answer is presented as feedback for 750 msec.



Predictors

For categorical judgments, the predictor measures the Euclidean distance of the stimulus from the category boundary in face-space (DIST). I did not collect data on reaction times for stimuli lying directly on the category boundary (DIST=0). In my analyses, therefore, I centered the DIST variable on the first meaningful value closest to the category boundary (DIST-1) so that the intercept parameter in the models could be interpreted directly.

Controls

There are four control variables: (a) subject's age in months, re-centered on 25 years = 300 months (AGE – 300), (b) subject gender (FEMALE, an indicator variable), (c) the number of years that the subject has owned a personal computer, re-centered on 10 years (COMP_YRS-10), and (d) the average number of hours per week the subject has spent using a computer during the past year, re-centered on 20 hours per week (COMP_HRS - 20). The centered values were chosen to make the parameters easier to interpret, assuming a typical subject is twenty-five years old, has owned a computer for a decade, and engages in moderate computer use between two and three hours per day on average throughout the week.

Data Analysis

RQ1: After some learning has occurred on the categorization task in human adults, do reaction times decrease systematically with stimulus distance from the category boundary?

Based on the learning behavior of the ANN simulation on the experimental task, I hypothesized that the relationship between DIST and ln(RT) would follow a linear level-1 model for individual human subjects, as follows:

$$\ln(RT_{ij}) = \pi_{0i} + \pi_{1i}(DIST_j - 1) + \varepsilon_{ij}$$

Where:
$\ln(RT_{ij})$ = Natural log of reaction time for subject *i* on stimulus *j* and is a linear function of the distance of item *j* from the category boundary ($DIST_j$)

π_{0i} = True natural log of reaction time for individual *i* when assessing a stimulus one unit away from the category boundary—that is, a stimulus with DIST=1 (level-1 intercept)

π_{1i} = True difference in the natural log of reaction time per unit of distance in stimulus space for individual *i* (level-1 slope)

ε_{ij} = Level-1 residual for individual *i* on stimulus *j*

σ_ε^2 = Level-1 residual variance across all occasions of measurement, for individual *i* in the population

At level-2, I specified a model to represent differences across individuals in the population in the level-1 intercept and slope, as follows:

$$\pi_{0i} = \gamma_{00} + \gamma_{01}(AGE_i\text{-}300) + \gamma_{02}(FEMALE_i) + \gamma_{03}(COMP\_YRS_i\text{-}10) + \gamma_{04}(COMP\_HRS_i\text{-}20) + \xi_{0i}$$

$$\pi_{1i} = \gamma_{10} + \gamma_{11}(AGE_i\text{-}300) + \gamma_{12}(FEMALE_i) + \gamma_{13}(COMP\_YRS_i\text{-}10) + \gamma_{14}(COMP\_HRS_i\text{-}20) + \xi_{1i}$$

Where:
γ_{00} = Population average true natural log of reaction time for a stimulus one unit away from the category boundary for a twenty-five year old male who has owned a personal computer for ten years and has used a computer twenty hours per week on average for the past calendar year

γ_{01} through γ_{04}: Difference in population average true natural log of reaction time for a stimulus one unit away from the category boundary between subjects one unit apart on the associated level-2 variable, controlling for the other level-2 variables. For example:

> γ_{01} = Difference in population average true natural log of reaction time for a stimulus one unit away from the category boundary for subjects one month apart in age, controlling for gender, years of computer ownership, and average weekly computer use over the past calendar year

ξ_{0i} = Level-2 residual on natural log of reaction time for individual *i* when assessing a stimulus one unit away from the category boundary, controlling for age, gender, years of computer ownership and average computer use over the last calendar year

σ_0^2 = Population residual variance of intercept π_{0i}, controlling for age, gender, years of computer ownership and average computer use over the last calendar year

γ_{10} = Population average true rate of change in natural log of reaction time per unit of distance in stimulus space for a twenty-five year old male who has owned a personal computer for ten years and has used a computer twenty hours per week on average for the past calendar year

γ_{11} through γ_{14}: Difference in population average true rate of change in natural log of reaction time per unit distance in stimulus space between subjects one unit apart on the associated level-2 variable, controlling for the other level-2 variables. For example:

> γ_{11} = Difference in population average true rate of change in natural log of reaction time per unit distance in stimulus space for subjects one month apart in age, controlling for gender, years of computer ownership, and average weekly computer use over the past calendar year

ξ_{1i} = Level-2 residual on average rate of change in natural log of reaction time for individual *i* when assessing a stimulus one unit away from the category boundary, controlling for age, gender, years of computer ownership and average computer use over the last calendar year

σ_1^2 = Population residual variance of rate of change π_{1i}, controlling for age, gender, years of computer ownership and average computer use over the last calendar year

My prediction is that the natural log of reaction time for people decreases linearly with increasing distance of the stimulus from the category boundary (recall Figure 3.6), operationalized here as the hypothesis that $\gamma_{10} < 0$. Therefore, to answer my research question I fit the hypothesized multilevel model to my experimental data and tested the null hypothesis H_0: $\gamma_{10} = 0$. A rejection of H_0 plus a negative sign on γ_{10} will be interpreted as evidence that the connectionist prediction (and by extension the connectionist model) is supported by the empirical evidence.

To address this research question, I fitted a hierarchy of nested multi-level regression models to the data based on the basic model specified above (see Appendix E for the full taxonomy of models). First, I fit the level-1 unconditional growth model as a

baseline for comparison. Next, I added all the level-2 predictors as a group (that is, I fit the level-2 model just as it is described above). As a group, the level-2 control variables did not significantly improve the fit of the model. Next, I examined the two-way interactions between the level-1 predictor DIST and each of the level-2 control variables. The only significant interaction at the $\alpha = .05$ significance level was the interaction between DIST and COMP_HRS. As a result, I removed the control variables AGE, FEMALE, and COMP_YRS to produce the final level-2 model[6], which includes the level-1 predictor DIST, the level-2 control COMP_HRS, and the interaction between them. The presence of the interaction term complicates the model, so I use prototypical plots to aid in the presentation of the findings. These plots illustrate the effect on ln(RT)xDIST as COMP_HRS varies from 5 hours (10^{th} percentile) to 60 hours (90^{th} percentile).

Results

Table 3.2 summarizes the results of fitting a nested hierarchy of multi-level regression models to the data for research question #1 (the main taxonomy is included in Appendix E). The leftmost column of the table lists the names of model components, organized into three groups: *fixed effects* (naming the parameters associated with the structural model components), *variance components* (naming the stochastic model components), and *goodness-of-fit* (including the -2 log likelihood goodness-of-fit statistic and the between- and within-person pseudo-R^2 statistics). Each row of the table thus contains a set of fitted values for a single model parameter across the different fitted

[6] I tested the stability of this final model by examining the effect of adding the control variables in various combinations. In particular, the DISTxFEMALE interaction was marginally significant on its own so I tried adding FEMALE to the final model as a main effect and also in the DISTxFEMALE interaction. As shown in Appendix E, none of the additional models I examined changed the final result qualitatively.

Table 3.2: Unconditional growth and final fitted linear multilevel models describing the relationship between ln(reaction time) in a dichotomous categorization task and the distance of the stimulus from the category boundary, controlling for subject's computer experience (average hours/week) and the interaction between stimulus distance and computer experience (subjects=48, observations=959).

	Model	
	Unconditional Growth	**Final Model**
Fixed Effects		
Intercept	6.7285****	6.7033****
	(0.03396)	(0.04109)
(DIST-1)	-0.0845****	-0.07093****
	(0.009748)	(0.01178)
(COMP_HRS-20)		0.002023
		(0.001865)
(DIST-1)*(COMP_HRS-20)		-0.00109*
		(0.000534)
Variance Components		
σ_ε^2	0.09112****	0.0907****
σ_0^2	0.04624****	0.04614****
σ_1^2	0.000	0.000
Goodness-of-fit		
pseudo-R_ε^2	0.076	0.080
pseudo-R_0^2		0.002
pseudo-R_1^2		0.000
-2LL	528.9	524.7

Key: ~ p<.10; * p<.05; ** p<.01; *** p<.001; **** p<.0001

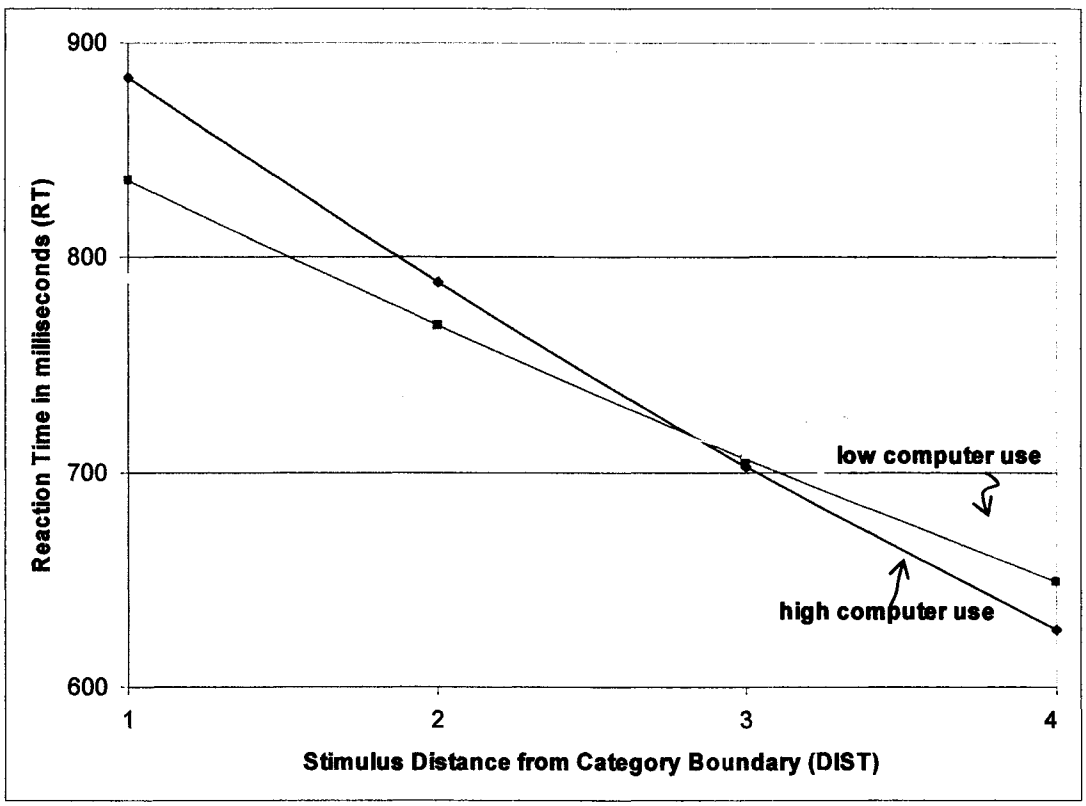
models. For the fixed effects, in parentheses below each parameter estimate is the associated standard error.

Two fitted models are represented in the remaining columns. The middle column describes the *unconditional growth* model, which includes the level-1 intercept and the main predictor DIST, both of which are statistically significant ($p < .0001$). The rightmost column describes the *final model*, which is the model I found that explained the most variance in ln(RT) with the least number of predictor variables. In addition to the level-1 intercept and main predictor, the final model also includes a statistically significant interaction between DIST and COMP_HRS ($p < .05$). The values of pseudo-R_ε^2 in Table 3.2 suggest that about 7.6% of the within-person variation in ln(RT) is explained by the main level-1 predictor DIST (pseudo-$R_\varepsilon^2 = .076$ for the unconditional growth model compared to the unconditional means model, shown in the first column of the table in Appendix E) and that this does not change much with the addition of the level-2 control variables (pseudo-$R_\varepsilon^2 = .080$ for the final model), as expected. Approximately 0.2% of the between-person variation in the level-1 intercept is explained with the addition of the level-2 predictor COMP_HRS and its interaction with DIST in the final model compared to the unconditional growth model (pseudo-$R_0^2 = .002$). The level-2 variance component associated with the true rate of change ($\sigma_1^2 = .000$) is not statistically significant in the unconditional growth model, suggesting that there is no residual variance in true rate of change to be explained with the addition of level-2 parameters (consequently, pseudo-$R_1^2 = .000$, suggesting no additional variance in true rate of change is explained by the addition of the level-2 predictors).

The final model includes a statistically significant interaction between the main predictor DIST and the control variable COMP_HRS ($p < 0.05$). The effect of the main predictor DIST on the outcome ln(RT) varies according to the value of COMP_HRS. The fitted slope of this interaction is -.00109 which means that on average in this sample of subjects, each additional hour of weekly computer use is associated with a difference of -.00109 in ln(RT) between stimuli located one unit of distance apart (moving perpendicular to and away from the category boundary). In other words, more computer use is associated on average with a greater decrease in reaction time between stimuli near the category boundary and those far away. The main effect of DIST (slope = -.07093, $p < .0001$) is the average difference in ln(RT) for two successively distant stimuli for a subject who spent an average of 20 hours per week using a computer in the past calendar year (COMP_HRS = 20).

The statistical interaction complicates the model interpretation because its presence means there is no single effect of the question predictor. To facilitate the presentation, therefore, I have chosen to display the findings as a set of fitted relationships between the untransformed outcome and the question predictor for several substantively interesting values of the control variable (Figure 3.10). On the horizontal axis I plot the distance of a stimulus from the category boundary. On the vertical axis I plot the predicted reaction time in response to the stimulus. Fitted curves of the untransformed outcome variable RT vs. the main predictor DIST are shown in Figure 3.10 for three representative values of COMP_HRS: 10^{th} percentile (5 hrs.), average (32.4375 hrs.), and 90^{th} percentile (60 hrs.). The relationship is curvilinear (taking the

Figure 3.10: Predicted reaction times as a function of distance from the category boundary by average number of hours of computer use per week



form of a decaying exponential) although the curvature is very slight[7]. The effect of computer experience is evident in the different slopes of the RTxDIST curves for different values of COMP_HRS. As the graph shows, more computer experience is associated with steeper reaction time curves. In other words, regular computer use seems to have a differential effect on different stimuli as a function of distance from the category boundary.

For the prototypical "low computer use" subject (COMP_HRS = 5), the average reaction time in response to stimuli close to the category boundary (DIST=1) is about 790ms, and reaction times decrease on average for stimuli further away. The difference in reaction times between stimuli close to the category boundary (DIST=1) compared to those furthest away (DIST=4) is about 120ms on average. For the prototypical "high computer use" subject (COMP_HRS = 60), the average reaction time for stimuli closest to the category boundary is about 880ms. This prototypical subject responds more slowly on average to stimuli near the category boundary than the "low computer use" subject. However, the difference in average reaction times for stimuli closest to and furthest away from the category boundary for this prototypical subject is more than twice that of the "low computer use" subject (257ms). In summary, the prototypical subject with high computer use does not respond uniformly faster than the subject with low computer use on all stimuli (as might be expected). He responds faster on average only to the stimuli furthest away from the category boundary. Instead, the pattern of results suggests that the

[7] Although the curvature is so slight that it is difficult to see upon visual inspection of Figure 3.10, the level-1 residuals are less heteroskedastic and more normal in the final model with the transformed output (ln(RT)) compared to a model with the same predictors but untransformed output (RT). In the final model reported here, the level-1 residuals still have a heavy upper tail, but otherwise they look reasonably homoskedastic and normal (see Appendix F).

high computer use subject's reaction times are more sensitive to stimulus distance from the category boundary.

Prediction #2: Perception of Visual Similarity Changes as a Function of Category Learning

Procedures

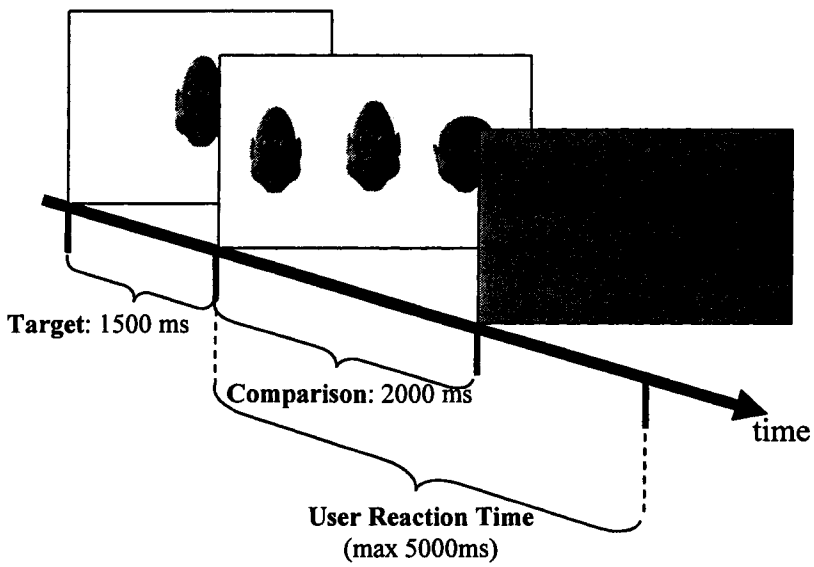
At the beginning of the experiment and after each block of one hundred categorical trials, I inserted a special block of twenty-two "similarity" trials to probe the subject's internal representation of the task similarity structure over time, as learning proceeds (Figure 3.11). On each similarity trial, I selected a target stimulus from one category (Gorf or Dimp) and a comparison stimulus from each category. The comparison stimuli were selected so that they are equidistant from the target stimulus in face-space. On each similarity trial, the target stimulus is presented to the subject briefly (1.5 sec), and then two comparison stimuli appear simultaneously to the right and left (for an additional 2 sec). The subject is instructed to press the arrow button (right or left) pointing toward the picture that looks most like the one in the middle. If a subject does not respond after 5 seconds, the trial times out and is marked as "no response."

Measures

Outcomes

In addressing RQ#2, I tested whether perceptions of stimulus similarity change as learning proceeds. On each similarity trial, I therefore recorded the subject's similarity judgment (SIM). Similarity judgments occur in blocks of 22 trials, on each of which the subject must decide which of two comparison stimuli (OBJ1 or OBJ2) is most similar to a target stimulus. OBJ1 is from the same category as the target, while OBJ2 is from the opposite category. If the subject selects OBJ1 then SIM = 1, otherwise SIM = 0. For

Figure 3.11: Structure of a similarity judgment trial. A face is presented for 1.5sec and then comparison faces are placed on either side of it for 2.0sec. Subjects respond by pressing an arrow pointing toward the face that looks most like the center face.



data analysis, I took the average value of SIM for each block of 22 trials, which is the fraction of trials on which OBJ1 was deemed most similar to the target (FR_SAME):

$$FR\_SAME = \sum_{k=1}^{22} SIM_k / 22$$

I treated FR_SAME as my outcome.

Predictors

For similarity trials, the predictor is TIME, recorded as the similarity block number (TIME=0 is the similarity block at the start of training, before any category learning has occurred, and TIME=4 is the last similarity block after the category training ends; the others are evenly spaced between blocks of categorical trials).

Controls

The same four control variables are used in this analysis as for the first prediction, centered on the same values (AGE-300, FEMALE, COMP_YRS-10, and COMP_HRS-20).

Data Analysis

RQ2: As human learning proceeds, do stimuli within a category come to be judged as increasingly similar to one another and less similar to stimuli in the alternate category?

To address RQ#2, I represented the relationship between FR_SAME and TIME by specifying a logistic level-1 model for individual subjects, as follows:

$$FR\_SAME_{ij} = \frac{1}{1 + \exp[-(\pi_{0i} + \pi_{1i} * TIME_{ij})]} + \varepsilon_{ij}$$

Where:
FR_SAME_{ij} is a continuous variable representing the fraction of trials on which the within-category stimulus, OBJ1, is deemed more similar to the target stimulus than is the cross-category stimulus, OBJ2, for individual *i* on trial *j*; it is a logistic function of TIME[8]

[8] Note that FR_SAME is a continuous variable created by averaging over blocks of 22 individual binary responses for each subject on each measurement occasion. Logistic models are usually created using the

π_{0i} determines the intercept, which is the true fraction of time OBJ1 is deemed more similar to the target stimulus on trial 0 (that is, before training) for individual *i*

π_{1i} determines the slope at midpoint, which is the true rate of change in FR_SAME for individual *i* at mid-point

ε_{ij} is the level-1 residual for individual *i* on trial *j*

σ_ε^2 is the level-1 residual variance across all occasions of measurement, for individual *i* in the population

At level-2, I specified a model to represent differences across individuals in the population in the level-1 parameters, as follows:

$$\pi_{0i} = \gamma_{00} + \gamma_{01}(AGE_i\text{-}300) + \gamma_{02}(FEMALE_i) + \gamma_{03}(COMP\_YRS_i\text{-}10) + \gamma_{04}(COMP\_HRS_i\text{-}10) + \xi_{0i}$$

$$\pi_{1i} = \gamma_{10} + \gamma_{11}(AGE_i\text{-}300) + \gamma_{12}(FEMALE_i) + \gamma_{13}(COMP\_YRS_i\text{-}10) + \gamma_{14}(COMP\_HRS_i\text{-}20) + \xi_{1i}$$

Where:

γ_{00} = Population average true value of π_{0i}, which determines the level-1 intercept, for a twenty-five year old male who has owned a personal computer for ten years and has used a computer twenty hours per week on average for the past calendar year

γ_{01} through γ_{04}: Difference in population average true value of π_{0i} between subjects one unit apart on the associated level-2 variable, controlling for the other level-2 variables

ξ_{0i} = Level-2 residual on value of π_{0i} for individual *i*, controlling for age, gender, years of computer ownership and average computer use over the last calendar year

σ_0^2 = Population residual variance of π_{0i}, controlling for age, gender, years of computer ownership and average computer use over the last calendar year

γ_{10} = Population average true rate of change of π_{1i} (which determines the rate of change of FR_SAME at midpoint) per block of trials for a twenty-five year old male who has owned a personal computer for ten years and has used a computer twenty hours per week on average for the past calendar year

γ_{11} through γ_{14}: Difference in population average true rate of change of π_{1i} per unit time (block of 100 categorical trials) between subjects one unit apart on the associated level-2 variable, controlling for the other level-2 variables. For example:

> γ_{11} = Difference in population average true rate of change of π_{1i} per unit time (block of 100 categorical trials) for subjects one month apart in age,

---

raw binary data. However, SAS V8 does not seem to support logistic nonlinear mixed models using dichotomous data at the observation level while also allowing for residuals at two different levels (observation level and individual level, in this case). I therefore had to first convert the binary data into probabilities to create a logistic nonlinear mixed model.

controlling for gender, years of computer ownership, and average weekly computer use over the past calendar year

ξ_{1i} = Level-2 residual on average rate of change of π_{1i} per block of trials for individual *i*, controlling for age, gender, years of computer ownership and average computer use over the last calendar year

σ_1^2 = Population residual variance of π_{1i}, controlling for age, gender, years of computer ownership and average computer use over the last calendar year

My prediction is that FR_SAME will increase over time on average, operationalized here as the hypothesis that $\gamma_{10} > 0$. To answer my research question, I fit the multilevel model to data and tested the null hypothesis H_0: $\gamma_{10} = 0$. A rejection of H_0 combined with a positive sign on γ_{10} will be interpreted as evidence supporting the connectionist model.

To address this research question, I fitted a hierarchy of nested multi-level regression models to the data based on the basic model specified above. First, I fit the level-1 unconditional growth model as a baseline for comparison. Next, I added each control variable as a main effect and simultaneously as a two-way interaction with the level-1 predictor TIME. The only significant effect at the $\alpha = .05$ significance level was the TIMExAGE interaction[9]. Therefore, the final model includes the level-1 predictor TIME, the level-2 control AGE, and the interaction between the two (TIMExAGE). The presence of the interaction term complicates the model, so I use prototypical plots to aid in the presentation of the findings.

Results

Table 3.3 summarizes the results of fitting a nested hierarchy of multi-level logistic models to the data for research question #2 (see Appendix G for the main

[9] The main effect of COMP_HRS was marginally significant, but this effect disappeared when both COMP_HRS and AGE were included in the model, suggesting the two predictors share variance (which is consistent with the Pearson correlation coefficient between COMP_HRS and AGE, which is $\rho = -.48$).

Table 3.3: Unconditional and final fitted logistic models describing the relationship between fraction of within-category pairs selected in a visual similarity judgment task and time (while learning was taking place) controlling for subject's age and the interaction between time and age (subjects=48, observations=240).

	Model	
	Unconditional Growth	**Final Model**
Fixed Effects		
Intercept	-0.04474 (0.05229)	-0.1468~ (0.07756)
TIME	0.1192** (0.03479)	0.2041*** (0.05102)
(AGE-300)		0.000599~ (0.000344)
TIME*(AGE-300)		-0.0005* (0.000223)
Variance Components		
σ_ε^2	0.008634****	0.008616****
σ_0^2	0.04355	0.03559
σ_1^2	0.04049**	0.03545**
Goodness-of-fit		
pseudo-R_ε^2	0.403	0.404
pseudo-R_0^2		0.183
pseudo-R_1^2		0.124
-2LL	-338.3	-344.3

Key: ~ p<.1; * p<.05; ** p<.01; *** p<.001; **** p<.0001

taxonomy). Table 3.3 is organized in the same way as Table 3.2; only the parameter names have changed.

The middle column describes the unconditional growth model, which includes the level-1 intercept (which is not significant at the .05 level) and the main predictor TIME ($p < .001$). The rightmost column describes the final model, which is the model I found that explained the most variance in FR_SAME with the least number of predictor variables. In addition to the level-1 intercept and main predictor, the final model also includes a statistically significant interaction between TIME and AGE ($p < .05$). The values of pseudo-R_ε^2 in Table 3.3 suggest that about 40.3% of the within-person variation in FR_SAME is explained by the main level-1 predictor TIME (pseudo-$R_\varepsilon^2 = .403$ for the unconditional growth model compared to the unconditional means model, shown in the first column of the table in Appendix G) and that this does not change with the addition of the level-2 control variables (pseudo-$R_\varepsilon^2 = .404$ for the final model), as expected. Approximately 18.3% of the between-person variation in the level-1 parameter that determines the intercept and 12.4% of the between-person variation in the level-1 parameter that determines the slope at midpoint is explained with the addition of the level-2 predictor AGE and its interaction with TIME in the final model compared to the unconditional growth model (pseudo-$R_0^2 = .183$ and pseudo-$R_1^2 = .124$, respectively).

The final model includes a statistically significant interaction between the main level-1 predictor TIME and the level-2 control variable AGE ($p < 0.05$). This means that there is no single main effect of the primary predictor TIME on the outcome FR_SAME—this effect varies according to the value of AGE. The fitted slope of this interaction is -.0005 which means that on average in this sample of subjects, each

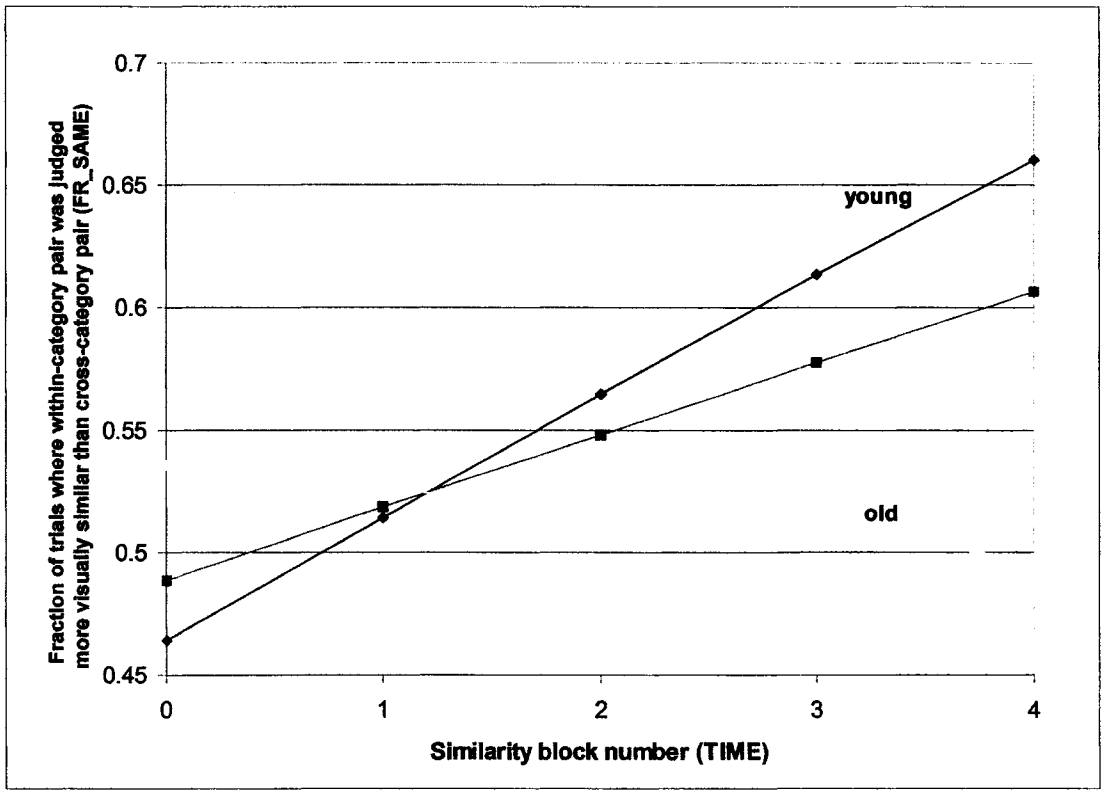
additional month of age is associated with a difference of -.0005 in the parameter that determines the slope of FR_SAME as a function of TIME. In other words, greater age is associated on average with less positive (or more negative) differences in FR_SAME at two successive sample intervals. The parameter determining the logistic intercept is not statistically significant in either model, which suggests that the fitted probability before training (at TIME=0) is not significantly different from 0.5 in either model.

The statistical interaction complicates the model interpretation because its presence means there is no single effect of the question predictor. To facilitate the presentation, therefore, I have chosen to display the findings as a set of fitted relationships between the outcome and the question predictor for several substantively interesting values of the control variable (Figure 3.12). On the horizontal axis I plot TIME. On the vertical axis I plot FR_SAME, which is the fraction of similarity judgment trials presented in that block on which the subject selected the within-category pair of stimuli as being more visually similar than the cross-category pair. Fitted curves of the outcome variable FR_SAME vs. the main predictor TIME are shown in Figure 3.12 for three representative values of AGE: 10^{th} percentile (25 years; 3 months), average (39 years; 2 months), and 90^{th} percentile (65 years; 6 months). The relationship is technically curvilinear (logistic[10]), but the data only range over a very linear portion of the curve.

The effect of age is evident in the different slopes of the FR_SAMExTIME curves for different values of AGE; as the graph shows, greater age is associated with decreased slopes. The prototypical "young" subject exhibited the predicted pattern, in which the

[10] The only assumption on the logistic model is that the sample log-odds of the outcome (FR_SAME) is a linear function of the predictor (TIME). This assumption seems to hold in this case (see graph in Appendix H of the log odds of FR_SAME vs. TIME).

Figure 3.12: Fitted trajectories of change during learning, for prototypical subjects, in the fraction of trials on which a same-category pair was identified as being more similar than a cross-category pair in a visual similarity judgment task



probability of selecting a within-category pair of stimuli as being more visually similar than an equally distant cross-category pair increases over time as learning of the categories proceeds. The prototypical "old" subject, on the other hand, did not show this pattern; instead of increasing (as predicted), the probability of selecting a within-category pair actually decreased slightly as the category training proceeded.

Discussion

The empirical results are consistent with both main ANN predictions. In the first case, reaction times are significantly longer for stimuli closer to the category boundary and shorter for stimuli further away, as predicted. Moreover, the use of a logarithmic transform on the outcome improves the model fit and the distribution of residuals compared to a linear model, which is consistent with the prediction about the form of the relationship (lower half of a negative logistic, or "squashed-S" function). The curves in Figure 3.10 are only slightly curvilinear, but this is not surprising given the relatively short learning time in the experiment compared to the simulation (400 trials for people compared to 5000 trials for the ANN). In particular, if the people were trained until they mastered all the stimuli as the ANN was, then I would expect (based on the first ANN model prediction) the curvilinearity of the reaction time curves to be more pronounced. This prediction could be investigated further by conducting an experiment in which the training continued until the stimuli were mastered or over-learned. It might even be possible to examine individual response curves in that case, since presumably the representations underlying the task performance would become much more stable at that point than they are early in the learning process.

Second, residual analysis indicates that the residual distribution has a long positive tail, but that the distribution looks more normal when the reaction time data is log transformed. The presence of a long tail was not predicted but is not surprising, given that reaction time data exhibit a floor effect (there is a limit on how quickly someone can respond, but no limit on how slowly she can respond).

One possible explanation for what is going on is that there could be two different underlying processes at work. In my experiment, the stimulus is presented for a short duration (400 msec), followed by a blank screen, during which time the subject is supposed to respond by categorizing the stimulus. Research on visual processing suggests that it takes a fraction of a second to identify the stimulus (Welford, 1980), and that visual stimuli are then maintained in a visual memory buffer for a while after the stimulus disappears. Estimates of this duration vary, but most researchers estimate this time to be under a second (Averbach & Sperling, 1961; Kosslyn, 1994; Sperling, 1960). It is possible that responses within the period during which the stimulus is either present or maintained in the visual buffer (or other short term memory buffer) will follow one distribution, and any responses made after the stimulus has leaked away engage different decision processes and therefore are drawn from a different distribution. It is plausible that people with less computer experience (who also tend to be the older subjects) would take longer to respond, either because they are less used to this kind of limited-time interaction, or because their unfamiliarity with computer interfaces and keyboards imposes a greater cognitive load on them compared to the more computer experienced subjects, which could contribute to a systematically slower response pattern. This slower response pattern would tend to push their response delays more often beyond the duration

of stimulus maintenance in the visual buffer, producing the observed pattern in the data. This effect could be further investigated or perhaps reduced by lengthening presentation time of the stimulus, training to complete mastery of the task, and/or providing more extensive initial training to prepare all subjects for the perhaps unfamiliar requirement of responding rapidly to a visual stimulus.

In the case of the second prediction, the fraction of same-category pairs chosen in the similarity judgment task increases on average over time with learning across the sample, as predicted (although the fitted curves for the oldest subjects in the sample became flat or even slightly negative). In addition, the logistic intercept is not significantly different from a probability of 0.5 (at TIME=0, which is before any training) after controlling for subject age and the interaction between age and time. In other words, subjects were just as likely on average to judge cross-category pairs of stimuli more visually similar than same-category pairs before any training on the underlying categories (controlling for age and the interaction between age and time). This finding is consistent with the assumption that these stimuli and categories are novel and have no intrinsic categorical structure (that is, there is no *a priori* bias to cluster particular groups of faces together).

For the similarity judgment task, the ANN predicted that the growth curves would be logistic between a probability of 0.5 before training and a probability of 1.0 after the task was mastered (note that this deviates from the more common scenario where the logistic relationship is assumed to range from 0.0 to 1.0). In the fitted model, the insignificance of the intercept parameter and positive sign of the growth parameter are consistent with this prediction, but the fitted curves in Figure 3.12 look more linear than

might be expected based on the prediction. One possible factor contributing to the linearity of the curves in this region of the trajectory is the functional form used for the statistical model. I was unable to fit the data with the ideal logistic functional form (which would have estimated the lower and/or upper asymptotes of the curve, depending on the sign of its growth parameter), in part because the data only cover part of the growth trajectory and in part because such models tend to be very unstable with even moderately noisy data. I used instead the familiar logistic function ranging from 0.0 to 1.0 to model the data in order to estimate the direction of growth and test the significance of the intercept, which is not significantly different from 0.5, controlling for age and the age-time interaction. A logistic function ranging from 0.0 to 1.0 and having an intercept of 0.5 is necessarily very linear in an interval around TIME=0. I was therefore unable to generate any evidence one way or the other on the shape of the growth trajectory in that region for this task for comparison to the ANN prediction.

The form of the interaction between time and age in this analysis suggests that on average, younger people tended to have a higher rate of increase in probability of selecting the same-category pairs in the similarity judgment task as time progressed compared to the older people. This could be an effect of age on learning the categories, and if so this effect could either operate directly (for example, due to the effects of age on neurology) or indirectly (for example, because older people had a harder time learning under these conditions with time-limited exposure, or because of increased cognitive load due to unfamiliarity with computer interfaces and input devices). This issue could be investigated further and possibly ameliorated in future studies by looking at longer training periods and/or increasing the stimulus presentation duration. If all subjects were

trained to mastery, for instance, based on the ANN prediction I would expect the older subjects to follow the same general trajectory as younger subjects on average but perhaps shifted in time and/or with a slower growth rate.

Implications

The main purpose of this study was to explore how computational modeling frameworks (such as the connectionist ANN framework) can be exposed to falsification through formal hypothesis tests involving empirical data. I began with the hypothesis that human brains employ a CNDR neural mechanism in learning novel categories. I used a computational model (ANN) embodying this neural mechanism to generate two behavioral predictions. I tested the behavioral predictions following from the neural hypothesis using human learning data and quantitative methods (multi-level regression models). The major findings are all consistent with the predictions, which means that we cannot reject the human-CNDR hypothesis based on this study.

A secondary purpose of this study was to demonstrate the feasibility of using ANNs in this particular way as part of a scientific method for researching causal brain-behavior relationships. Based on the results of this study, I would say this approach is not only feasible, but quite promising. The CNDR hypothesis is based on observations about the mammalian nervous system, the logic connecting the biological mechanism to the ANN is made explicit, the effects of the CNDR mechanism at the "neural" level in the ANN are traced to specific "behavioral" patterns caused by it in the ANN, and there is an explicit logic linking model behavior to human behavior (through the formal specification of a mapping from the face stimuli to their coordinates in face-space).

Critics can always take issue with any step of this argument, of course, but this is precisely the point; I have attempted to explicate and justify the logic at every step along the way to expose the process to refinement or refutation grounded in evidence from analytic and empirical methods in order to move closer to an accurate theory of brain-behavior links.

Moreover, the novel experimental paradigm and the novel application of statistical analysis to link ANN model predictions to human data demonstrates one way to fill an important gap in the domain of brain-behavior research. Specifically, much more use could be made of computational models like the ANNs used here if they were part of a systematic, explicit scientific research process instead of being limited to theoretical exploration, hypothesis generation, and thought experiments, as is far more common (although these are also very valuable applications of the models). The design described in this study provides one experimental paradigm and one set of methods that can be applied together as I have shown to move closer to that goal.

Finally, this experiment serves as a concrete case study demonstrating how this approach can be realized in practice, and how it fits into the larger research frameworks (the basic brain-behavior research method and the applied educational neuroscience research framework) that are the main focus of this dissertation. Each component of this design (experimental paradigm, CNDR neural mechanism, ANN model, predictions, statistical modeling applied to cognitive and microgenetic data) could potentially be applied to investigate many other questions, some generated during the course of this study and others far removed. For example, the methods could be adapted to investigate how different strategies used by subjects relate to the reaction time data, to investigate the

relationship between reaction time data and response accuracy, to study whether subjects' subjective reports of their behavior predict their actual behavioral data, and to explore whether the human and simulated data can be used to understand more about how they relate to one another (for example, how a unit of simulation time corresponds with a unit of human time).

The educational implications of this study are necessarily speculative and would require independent evaluation if used to inform practice, as I argued in the dissertation introduction. There are a number of interesting insights and possibilities, however. For example, what appears to be the "same" task given to a group of people from the perspective of the teacher (or experimenter in this case) is not necessarily the same task from the perspective of the students (or subjects). This general point is intuitively obvious, perhaps, to anyone who has ever taught, but the methods employed in this study could potentially enable researchers and teachers to sharpen and formalize their intuitions on this point to manage this aspect of teaching more effectively.

In the present study, I thought initially that the parameters of the task were transparent enough that virtually everyone would approach it in roughly the same way—in particular, using more or less the same or equivalent visual information as the basis for their categorization strategies (since I designed the stimuli systematically by varying two dimensions). This was not the case, however, as I discovered in post-experimental interviews (which I only allude to here, since these data were not formally part of this initial study but were collected primarily to help contextualize the quantitative data and inform revisions to the experimental design for future studies—see Appendix I for the post-experimental questionnaire).

From the ANN perspective, this learning problem always involves two integrated dimensions of information (the face height and head shape). From the human perspective, people evidently experienced it in a variety of ways. Some reported seeing it as a one-dimensional problem (for example, using only head shape to make their categorical decisions), some described a two-dimensional strategy where two "rules" were applied sequentially (for example, "look at the head shape, and if that is not definitive then look at the expression"), and some described strategies that integrated two dimensions into a single decision rule like the ANN did (for example, using a complex "feature" of the faces such as emotional valence). In some cases, people used several rules (more than two), and in others people reported using one or more rules and memorizing the "exceptional" faces that they found difficult. This task clearly generated a rich set of response strategies.

The educational insight is that if a task this straightforward elicits such diverse strategies, then there are probably few tasks used in schools that are as well controlled as teachers might desire or expect. On the positive side, the diversity was not infinite. Most subjects noticed and used head size (or distance between the eyes), often in conjunction with one or more other salient (and relevant) features. Moreover, although subjects used many different labels for their strategies, some relied on essentially the same information drawn from a different source (for instance, forehead size covaries with the "scrunchiness" of the facial features and with the height of the chin), so these can be equated with one another. Some strategies are just variations on a single theme using different labels, such as strategies based on nice vs. mean faces, or approachable vs. avoidable faces, or female vs. male faces, or friendly vs. scary faces.

The task was specifically designed to be two dimensional. It should be possible, therefore, to organize all strategies into a set of categories based on whether they use one or two dimensional strategies. The two-dimensional strategies can then be further differentiated into sequential (multiple one-dimensional rules) or integrated (a single two-dimensional rule). The resulting categories can be differentiated still further to deal with strategies involving more than two rules and those explicitly identifying exceptions to any of the specified rules. This kind of approach would preserve the meaningful variability while providing tools for limiting and managing the complexity of strategy profiles. The experimental paradigm and task domains used in this experiment seem promising for investigating educational issues such as these more explicitly.

A more direct implication of the current study is insight into why a set of items that should be uniformly difficult (or easy) to learn might not be. In this study, for example, it might seem obvious *a priori* that all the faces should be equally difficult to learn, but the results of the experiment suggest that the difficulty of learning the correct category membership for a given face varies systematically with the distance of that face from the category boundary. In my mind, the interesting point highlighted by the current experimental design is that the item difficulties are not necessarily due to any intrinsic properties of the items themselves—instead, they emerge through an interaction between the intrinsic *relationships* among the items, the category structure imposed externally on the set of items, and properties of the learner's nervous system. Just recognizing the existence of this set of relationships could be the basis for a useful educational design principle, and the more we understand about these relationships in isolation and in relation to one another, the more powerful these design principles would be likely to be.

Another implication of this study derives from the second behavioral prediction, about the effect of learning on perception of similarity. Understanding this mechanism could help us to understand at least one aspect of how previous learning influences future learning. In other words, if one effect of learning a set of categories is to change systematically our perceptions of visual similarity, this would be a very interesting mechanism. It could be useful, for example, to help curriculum designers extend their thinking beyond the problem of what content should be included in a lesson to thinking also about how that content should be organized in time to capitalize on the cumulative effects of previous learning on students' perceptions of new material. The most obvious domains where the effect on visual perceptions might be useful are domains with a strong visual component, such as fine art or art history. If this effect of the CNDR mechanism operates in the visual modality, however, it very likely operates in the other sensory modalities as well, so the research program and any useful educational principles could very likely be generalized quickly to these other spheres.

Any study involving cognitive simulations inherits some limitations of those methods. In particular, it is often difficult to know what inferences are valid from the simulation to human cognition (Churchland, 1988; Gershenfeld, 1999; MacDonald & MacDonald, 1995). I believe that the methods described here represent an incremental advance in addressing this problem by providing a way to suppress the idiosyncratic details of simulated learning trajectories and foregrounding more abstract general characteristics of model behavior that can be compared to empirical data from people using statistical hypothesis testing. Whether this approach can be generalized to other paradigms and domains is a question for future study.

Since this is the first study of its kind, I have proposed a novel experimental task that is not clearly related to real educational content. I did this to maximize experimental control over many extraneous variables while I tested the proposed methods and hypotheses. This is common practice in learning experiments (see, for example, Bruner, Goodnow, & Austin, 1956), and I expect the results and conclusions can be generalized eventually to more relevant educational content.

At the very least, it should be possible to apply this experimental paradigm to more natural kinds of tasks—for example, to investigate the conceptual organization induced by a particular curriculum. In particular, the experimental task was specifically designed so that a dichotomous category structure (Gorf vs. Dimp) is being imposed upon a continuous underlying space of stimuli (the face-space, which varies continuously along two dimensions—head width and distance from mouth to eyes). The effects I investigated in the experiment arise from interactions between these two different "similarity structures"—one inherent and perceptual (the visual similarities between faces) and the other arbitrary and conceptual (the arbitrary category labels imposed on the stimuli).

Several educationally relevant task domains have a similar kind of structure, where continuous perceptual dimensions interact with more abstract or conceptual dichotomous categories. For example, in music theory the notes on a musical staff are arranged according to the continuous perceptual dimension of pitch—higher pitches are located higher on the staff. Imposed on this continuous dimension of pitch are discrete octave categories, in which, for instance, all of the "C" notes of different pitches are grouped together into a single category. The interaction of the continuous dimension of

pitch with the dichotomous dimension of octave category could have implications for learning music theory, and in particular could have implications for how it should be taught. The experimental paradigm in the present study could be used to investigate the structure of music theory knowledge induced by different curricula and teaching techniques. Similarly, the periodic table in chemistry is organized according to an underlying continuous dimension of atomic weight, upon which is imposed a set of discrete functional categories (noble gases, +1 valence, -1 valence, rare earth metals, etc.) depending on such attributes as the number of valence electrons available in the outer energy shell. In addition, categorical species distinctions in biology (e.g., dog, cat, human) are imposed on a more continuous space of genetic differences (every human has a unique genome). As these examples illustrate, many educationally relevant domains or sub-domains are structured like the experimental task at an abstract level and could therefore be studied using similar methods.

The present study suggests that one non-obvious source of complexity in knowledge domains such as these could be the interaction of the continuous and categorical organizing dimensions, which might make some individual items more difficult to learn than others even though the items appear on the surface to be highly uniform (for example, two elements next to each other on the periodic table). This effect could operate even in the most rudimentary learning scenarios, for example by making some elements of the periodic table more difficult to memorize and more prone to errors during recall than others.

Conclusions

Psychologists have used computer simulations in their research for decades. Some proponents of the approach have conceptualized it as an empirical (rather than experimental) paradigm having little use for standard statistical techniques because of its focus on the complex, history-dependent details of individual (not group) behavior (Newell & Simon, 1972). In my opinion, however, the practical utility of these models has been limited because researchers have paid insufficient attention to the question of how model predictions can be tested formally against empirical data in ways that expose the modeling paradigms themselves to falsification.

In this study, therefore, I have described methods for testing predictions from a connectionist model using empirical learning data from a sample of people. These methods leverage certain strengths of process models, such as their ability to capture individual learning trajectories (Newell & Simon, 1972), while making them accessible to standard statistical techniques. I expect that this methodology could be extended to other types of learning and other kinds of hypotheses.

In addition, the outcome of this study has implications for the larger neuroscience-education debate going on in the field. In particular, if human learning behavior is significantly shaped by similarity structure as predicted by the CNDR neural mechanism embedded in the connectionist model, then this represents a glimpse into a powerful and rather direct neurobiological constraint on observable learning behavior (that is, this would represent a concrete brain-behavior link), contradicting claims that it is not possible to link neurobiology directly to educationally relevant issues (Bruer, 1997). In one way or another, I hope this study will contribute incrementally to the larger

effort underway to establish a scientific basis for education grounded in studies of the brain and mind, both methodologically and substantively.

In my mind, the study described here is just the first small step into a very large research territory, with potentially far-reaching implications, many of which have educational relevance. For example, the ANN sensitivity to similarity structure has been linked theoretically to knowledge transfer (Connell, 2002; Plunkett & Elman, 1997), learning and development (Anderson, 1995; Elman et al., 1996; McLeod et al., 1998), affect and motivation (Fischer & Connell, 2003; Sutton & Barto, 1998), learning disabilities (Cohen, Sudhalter, Landon-Jimenez, & Keogh, 1993; Oliver et al., 2000; Plaut & Shallice, 1993), and the structure of internal representations (Quinn & Johnson, 1997). I expect that the kind of research proposed here will help us "reverse-engineer" such phenomena, thereby suggesting innovative, scientifically grounded design principles for effective educational interventions and assessments. In the next chapter, I discuss how the brain-behavior link investigated in this experimental study can be linked to educationally relevant theory. I illustrate the process with a concrete example linking the CNDR neural mechanism to behavioral patterns associated with knowledge transfer.

Chapter 4

Educational Implications: Neural Models of Knowledge Transfer

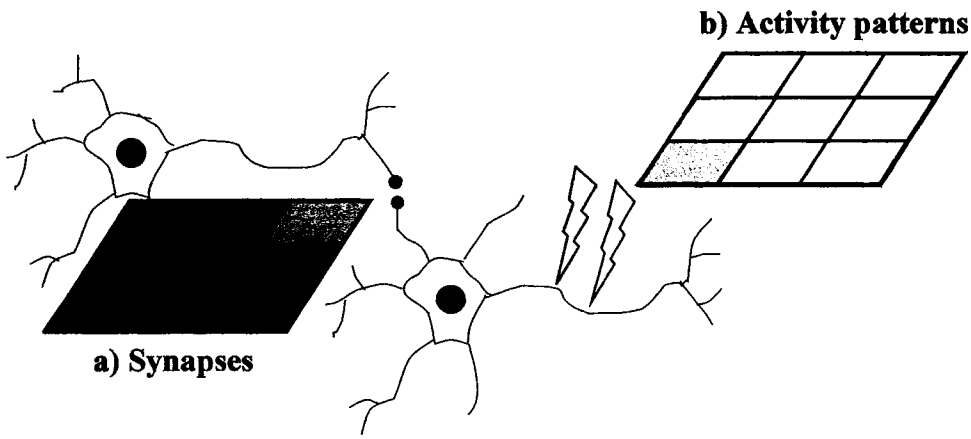
Introduction

Until recently, the domains of neuroscience and psychology evolved largely independently of one another (Bruer, 1999a). In particular, purely psychological and cognitive theories have rarely incorporated neurological constraints (but see James, 1890 for a notable early exception). This theoretical separation between brain and mind generally rests on the assumption that brain properties do not "show through" to the psychological and behavioral levels (e.g., Simon, 1992), which leads people to conclude that theories of mental function can be developed independently of theories of brain function (Marr, 1982). In this paper, I challenge the common assumption that psychological theories are or can be theoretically insulated from neural considerations. My argument is based on a comparative analysis illustrating how two mutually exclusive assumptions about brain mechanisms support qualitatively different theories of psychological and behavioral phenomena, using knowledge transfer as a concrete example.

Theoretical Background

For the purposes of this discussion, the only relevant neuroscience finding is the generic observation that the brain employs two distinct mechanisms to store knowledge: synapses and activity patterns (Figure 4.1). The biological details of how these mechanisms operate and how they differ from each other are interesting in their own right, but they would merely complicate without facilitating the present analysis. The

Figure 4.1: Two different mechanisms employed by the nervous system to store knowledge: a) synapses (stable patterns of physical connections between neurons); b) dynamic patterns of neural activity (e.g., spike trains)



interested reader can find information on this subject in any neuroscience textbook[1] (see, for example, Bear, Connors, & Paradiso, 1996; Kandel, Schwartz, & Jessell, 2000).

This simple fact about the nervous system suggests a very straightforward question: *How might the information encoded in synapses be related to the information encoded in activity patterns?* Logically speaking, there are a number of possible ways that two sets of representations could be related to one another. For example, the two different systems could simply be a means for the brain to double its information storage capacity (like adding an additional hard drive to a personal computer), in which case the content stored in the two sets of representations could be completely independent of each other (the same way computer files stored on two separate hard drives are independent of one another). This scenario is logically possible but not very plausible physiologically. Synaptic representations are comparatively durable but inflexible, while activity-based representations are flexible but short-lived. These qualitative differences between the two systems make it unlikely that they are interchangeable. More importantly, this possibility is not very interesting for present purposes because it does not have any obvious implications for psychological or behavioral phenomena. The interesting configurations of a two-mechanism information storage system (like the nervous system) are those wherein the two sets of representations are coordinated and therefore constrain each other in some way.

Two possible ways to coordinate two sets of distributed representations[2] are: 1) to make the two sets of representations contain the *same* information (i.e., make them copies of one another), or 2) to make the two sets of representations contain *different* (but

[1] In Chapter 2 I also give a brief overview of the two kinds of representations.
[2] See Chapter 2 for a more formal and thorough derivation of this argument.

interdependent) information. To illustrate these two possibilities, imagine you have collected data on shoe size and math achievement from a number of children (Table 4.1).

Table 4.1: Imaginary data on children's shoe sizes and performance on a math achievement test.

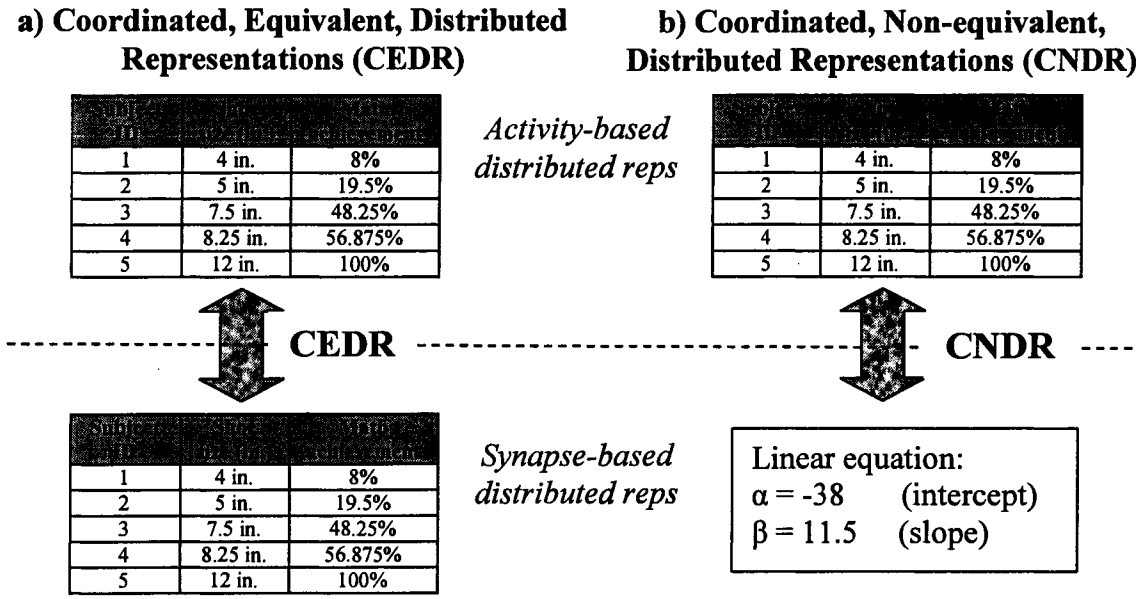
Subject ID	Shoe size (in.)	Math achievement
1	4 in.	8%
2	5 in.	19.5%
3	7.5 in.	48.25%
4	8.25 in.	56.875%
5	12 in.	100%

Using two sets of distributed representations, there are two distinct ways these data could be stored in an information processing system (whether it is a brain, a digital computer, or some other kind of device):

Coordinated, Equivalent, Distributed Representations (CEDR): In this case, the two sets of distributed representations contain the same information, although perhaps in different formats (see Figure 4.2a). For example, if the data in Table 4.1 were entered into a spreadsheet on a computer, the spreadsheet loaded into the computer's random access memory (or RAM, which is the computer's working memory) would be like the activity-based representation. If the file were then saved, an exact copy of the spreadsheet contents would be copied from RAM onto the hard disk (the computer's long-term memory). The file stored on the computer's hard disk would be analogous to the synapse-based representation. In this scenario, these two sets of representations contain identical information—both veridically store the set of data points listed in Table 4.1. These two sets of representations (the spreadsheet in RAM and the spreadsheet file on hard disk) are, for all intents and purposes, copies of one another.

Coordinated, Non-equivalent, Distributed Representations (CNDR): In this configuration, the two sets of distributed representations contain different (but interdependent) information. For example, note that the data in Table 4.1 exhibit a

Figure 4.2: Two ways to coordinate two sets of distributed representations: a) both sets contain the same information (CEDR), or b) the two sets contain different information (CNDR).



1	4 in.	8%
2	5 in.	19.5%
3	7.5 in.	48.25%
4	8.25 in.	56.875%
5	12 in.	100%

1	4 in.	8%
2	5 in.	19.5%
3	7.5 in.	48.25%
4	8.25 in.	56.875%
5	12 in.	100%

1	4 in.	8%
2	5 in.	19.5%
3	7.5 in.	48.25%
4	8.25 in.	56.875%
5	12 in.	100%

perfectly linear relationship. The contents of the table can be summarized without error using the linear equation:

$$\text{Math\_achievement} = -38 + 11.5 * \text{Shoe\_size}$$

This equation implicitly contains all of the information in Table 4.1. When we want to work with the shoe size and math achievement data (in the activity-based representations) we need access to the actual numbers themselves, but this does not mean that we have to *store* those numbers directly in the synaptic representations for direct recall. Instead of storing the table itself we could just as easily (if not more easily) store the linear equation parameters (intercept=-38 and slope=11.5) in the weight-based representations and use these to generate values of Math_achievement on demand, given particular values of Shoe_size[3]. These two sets of information (the slope and intercept on the one hand and the shoe size and math score pairs on the other hand) are clearly *coordinated* with one another (the equation parameters generate the numerical values and vice versa), but they are also obviously not *copies* of one another (for example, note that none of the individual numbers in the table bears any relationship to the values of the slope and intercept). This example illustrates concretely how the two sets of representations in a CNDR system can be coordinated with one another and yet can contain completely different information (Figure 4.2b).

---

[3] Note that I am not suggesting the linear relationship is stored explicitly in the system in such a way that the rule itself needs to be recalled to consciousness before it can be used. Instead, the neural circuitry would *embody* the generative equation. The person (if we think for a moment in terms of the CNDR mechanism operating in a human nervous system) would have no direct access to the rule itself; she would only have access to the number facts that it produces in the activation patterns. One implication of this fact is that the CEDR and CNDR mechanisms would not necessarily be distinguishable using self-report data since people would only have access to one of the two sets of representations (the activation patterns) in either case. Note, however, that a lack of conscious access to contents of synaptic representations does not necessarily imply that these representations can be treated independently of the activity-level content that is consciously accessible.

Research Question

In the previous section I described two different mechanisms whereby two sets of distributed representations could be coordinated: CEDR and CNDR. These two mechanisms represent mutually exclusive hypotheses about brain organization, based on the single neuroscience finding that the brain uses two distinct systems to store knowledge (synapses and activation patterns). These two hypotheses are the point of departure for the research question that is the main focus of this analysis: *Do different assumptions about the brain support qualitatively different theories of particular psychological and/or behavioral phenomena? If so, then how?*

In the following sections I argue that different assumptions at the neural level do, in fact, have differential implications for theories of psychological and behavioral phenomena. I ground the analysis concretely by showing how the two neural hypotheses lead to different theories of learning, recall, and knowledge transfer.

Analysis

I make two assumptions in the following analysis:

1) Synaptic representations are more durable than those encoded in activation patterns, but the synaptic representations are not directly accessible for processing (including not being available to conscious processes)

2) Activity-based representations are transient, but they can be processed flexibly and in some cases made accessible to consciousness

Knowledge Acquisition and Knowledge Application

At the most basic level, the two neural hypotheses imply different models of knowledge acquisition (learning) and recall. The CEDR mechanism implies a "storage"

or "container" model of knowledge representation (Figure 4.3), in which the senses provide a constant stream of information in the form of activation patterns, some of which are selectively "saved" into the synaptic representations for long term storage. The information stored in the synaptic representations is durable but cannot be manipulated directly. During the recall or application phase, therefore, knowledge must first be copied into the more flexible dynamic activation patterns (working memory) for recall and processing. The "container" metaphor is appropriate for this kind of mechanism (Figure 4.3b), where the knowledge being stored is the "object," the synaptic representations (long-term memory) constitute the "container" where these knowledge objects are stored, and the processes of learning and recall are analogous to the actions of placing objects into and removing objects from the container, respectively.

A concrete example is a spreadsheet being manipulated on a digital computer (Figure 4.3c). Each data element (e.g., each value of shoe size) is stored in its own spreadsheet cell; the cell contents are individual knowledge objects. The copy of the spreadsheet stored in RAM contains a set of objects that have been removed from the synaptic "container" for manipulation. When the spreadsheet is saved, the knowledge objects in the spreadsheet cells are copied directly to a file on the hard drive, which is analogous to placing the objects in the storage container.

The CNDR mechanism, in contrast, supports a "machine" or "generative" model of knowledge representation (Figures 4.4 and 4.5). Knowledge is not "stored" explicitly in this kind of system. Instead, related patterns of activation (inputs and outputs) are held in the working memory (activity patterns) providing an example of the kind of "machine" that is required at the synaptic level.

Figure 4.3: Learning and recall in a CEDR system. Three views on the CEDR mechanism: a) two sets of distributed representations (synaptic and activity-based) containing the same information; b) the "container" metaphor, in which knowledge objects are moved in and out of long-term storage; c) a real-world example of a CEDR mechanism is a spreadsheet on a computer. Learning, in this system, is a process of copying the contents of activity-based representations directly into synaptic representations for long-term storage. Stored knowledge is recalled by copying it back from synaptic representations into the activity-based representations for processing and application.

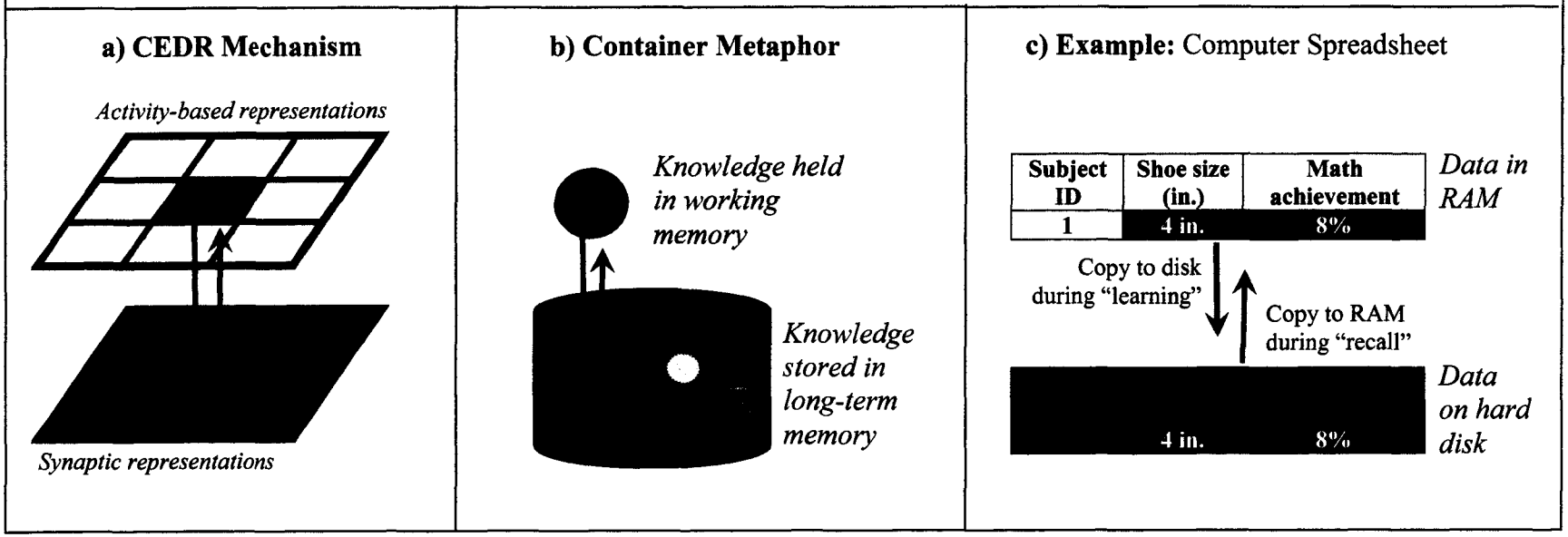


Figure 4.4: Learning in a CNDR system. Three views on the CNDR learning mechanism: a) two sets of distributed representations (synaptic and activity-based) containing different information; b) the "machine" metaphor, in which knowledge is stored implicitly in the form of a machine that converts input to outputs on demand; and c) a real-world example of a CNDR mechanism is a linear regression equation summarizing a relationship between two (or more) variables (inputs and outputs). Learning, in this system, is a process of using input-output pairs as prototypes to build a machine (neural circuit) embodying the input-output relationship implicitly.

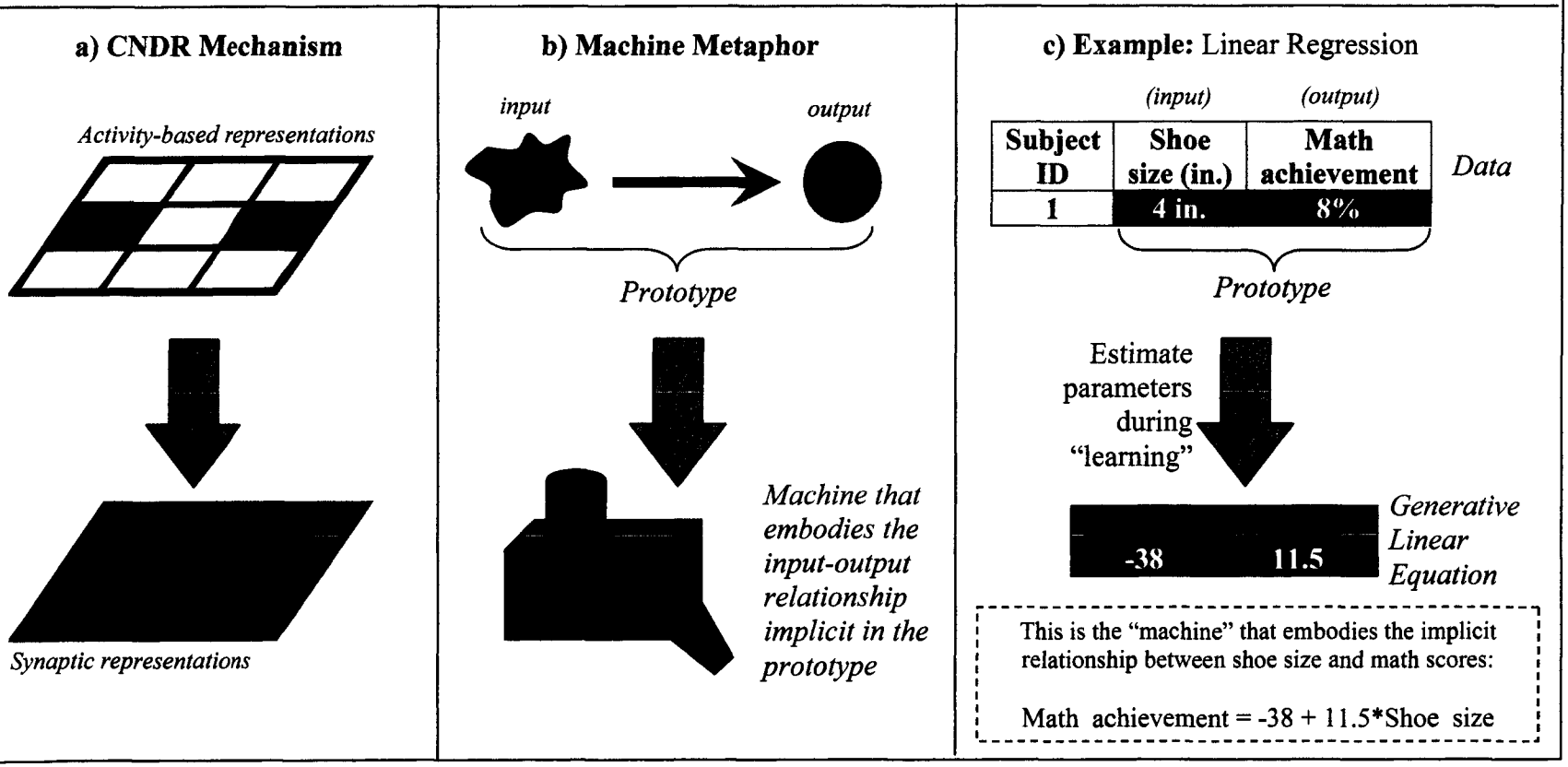
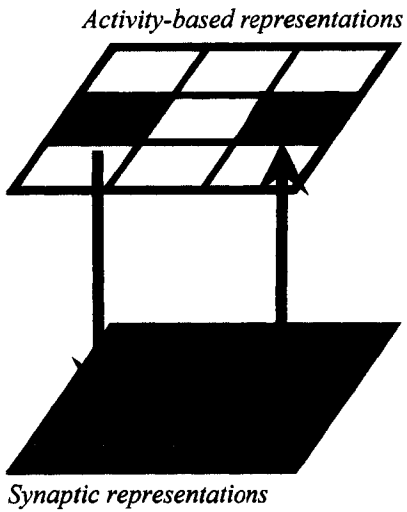
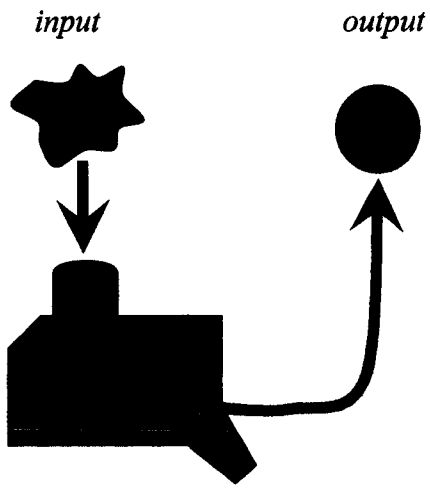


Figure 4.5: Recall in a CNDR system. Three views on the CNDR recall mechanism: a) two sets of distributed representations (synaptic and activity-based) containing different information; b) the "machine" metaphor, in which the appropriate output is generated anew each time an input is fed into the machine; and c) a real-world example of a CNDR recall mechanism is a linear regression equation to which an input has been applied to produce the associated output. Recall, in this system, is a process of feeding inputs to the appropriate neural circuit (machine) to re-generate the desired outputs (knowledge) on demand.

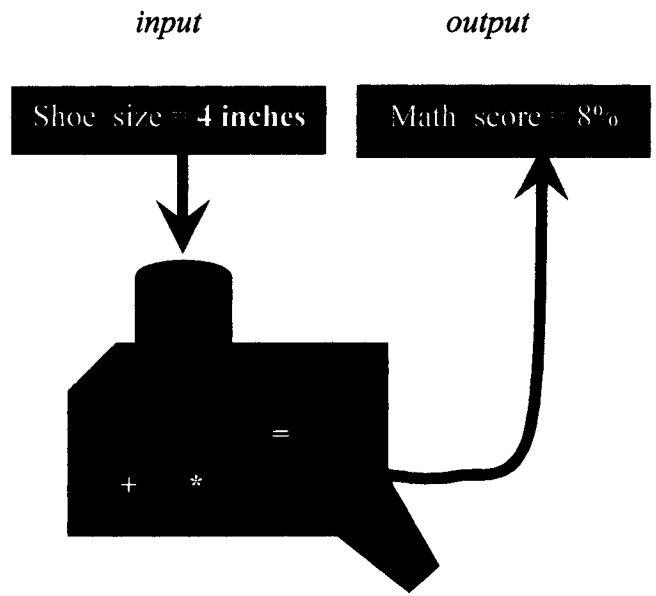
a) CNDR Mechanism



b) Machine Metaphor



c) Example: Linear Regression



By analogy, imagine an inventor who wants to go into business manufacturing widgets. He would approach an engineer with a sample (or description) of the raw material (e.g., molten plastic) and a prototype of the final product (e.g., a spherical widget) and ask the engineer to design a machine that converts the raw material into finished products like the prototype (Figure 4.4b). This scenario is analogous to the learning and recall model in the CNDR system. The raw materials correspond to inputs (e.g., sensory stimuli), the widgets correspond to outputs (e.g., motor responses), the process of machine design is the process of learning, and the machine itself is a neural circuit that is constructed by changing synaptic connections. Importantly, the product of learning in this model is not the explicit knowledge that is being represented (the widget)—in contrast to the CEDR / container model. Instead, the product of learning is a machine capable of generating specific knowledge (widgets) on demand, given appropriate inputs (raw materials). During each recall episode, the inputs are fed into the neural machine and the appropriate outputs are generated anew (Figure 4.5b).

As a concrete example of the CNDR mechanism in action, consider again the math achievement data set. In this model, the individual cells are not treated as independent knowledge objects as they were in the container model. Instead, each shoe size measure and its associated math achievement score are treated as a prototype input-output pair. The product of learning is a neural circuit (machine) that produces the appropriate math achievement score as its output (widget) when a shoe size measurement (raw material) is fed into it (Figure 4.4c). The machine, in this case, takes the form of an input-output relationship (equation) embodied directly in the neural tissue, produced by learning processes that change synaptic connections (equation parameters) the way a

machine design engineer would take standard components (gears, motors, etc.) and connect them together in specific novel configurations (analogous to changing synaptic connections) to produce a machine exhibiting the desired input-output behavior. The CNDR recall process in this example (Figure 4.5c) would involve feeding a shoe size measure as input (raw material) into the neural circuit (machine) to generate the associated math achievement score (widget). This example focuses on a single simple machine to simplify the discussion. In a CNDR system at the scale of the human brain, there would be many, many billions of such machines connected together into larger networks where one machine's inputs could come from the outputs of many other machines and its outputs could in turn potentially be fed as inputs into many other machines. The natural extension of the machine analogy in this case is to multi-stage processes in a factory utilizing many machines, or even to a series of factories involved in converting raw materials into final products.

Knowledge Transfer

As the discussion in the preceding section illustrates, the two neural hypotheses (CEDR and CNDR) clearly support qualitatively different theories of learning and recall. The question is whether these differences are isolated at the neural level, or whether they "show through" in some way to the levels of psychological and behavioral phenomena. To address this question, I examine the educationally relevant phenomenon of knowledge transfer to provide a concrete example illustrating how psychological- and behavioral-level theories of such phenomena are sensitive to underlying (explicit or implicit) neural assumptions.

Knowledge transfer is the process of applying knowledge learned in one context (for example, addition learned in a classroom setting) to a different context (for example, to calculate the total cost of a group of items in a store based on the prices of the individual items). Transfer has been a focus of psychological inquiry since at least the beginning of the twentieth century, because it is important to a wide range of phenomena in cognitive science generally and education specifically. In particular, knowledge transfer must happen for a student to apply almost anything learned in a classroom context to a real-world situation.

Classically, psychologists have defined knowledge transfer in terms of two main variables: distance (how far?) and amount (how much?). Transfer distance is typically defined in terms of a continuum from "near" to "far," depending on how different the context of learning is from the context of application. For example, learning to drive a car and then being able to drive a rental truck with little or no additional training is a case of near transfer because the two task contexts are quite similar. Learning to play chess and then applying chess principles (e.g., "material advantage" or "control of the center") to business situations (e.g., in planning a hostile takeover or selecting a site to locate a store, respectively) would be examples of far transfer (Salomon & Perkins, 1989).

Transfer amount is typically operationalized by assessing how much the knowledge acquired in the learning context facilitates performance in the application context. For example, a study might examine the extent to which learning a programming language facilitates performance on specific deductive reasoning tasks compared to a control condition in which no such training is provided.

Despite its importance and decades of focused attention, however, transfer continues to be a puzzling phenomenon. In particular, researchers have failed to produce widely applicable general principles about transfer (Fischer & Farrar, 1987), and there are many contradictory findings on transfer reported in the literature (Salomon & Perkins, 1989). Because of these difficulties, some application-minded researchers have advocated shifting the focus away from the theoretical questions of how transfer operates and onto more practical issues, such as the conditions and behaviors that facilitate transfer (Salomon & Perkins, 1989). In part, this recommendation is based on the observation that several fundamental questions about transfer have remained unresolved despite decades of focused research, including:

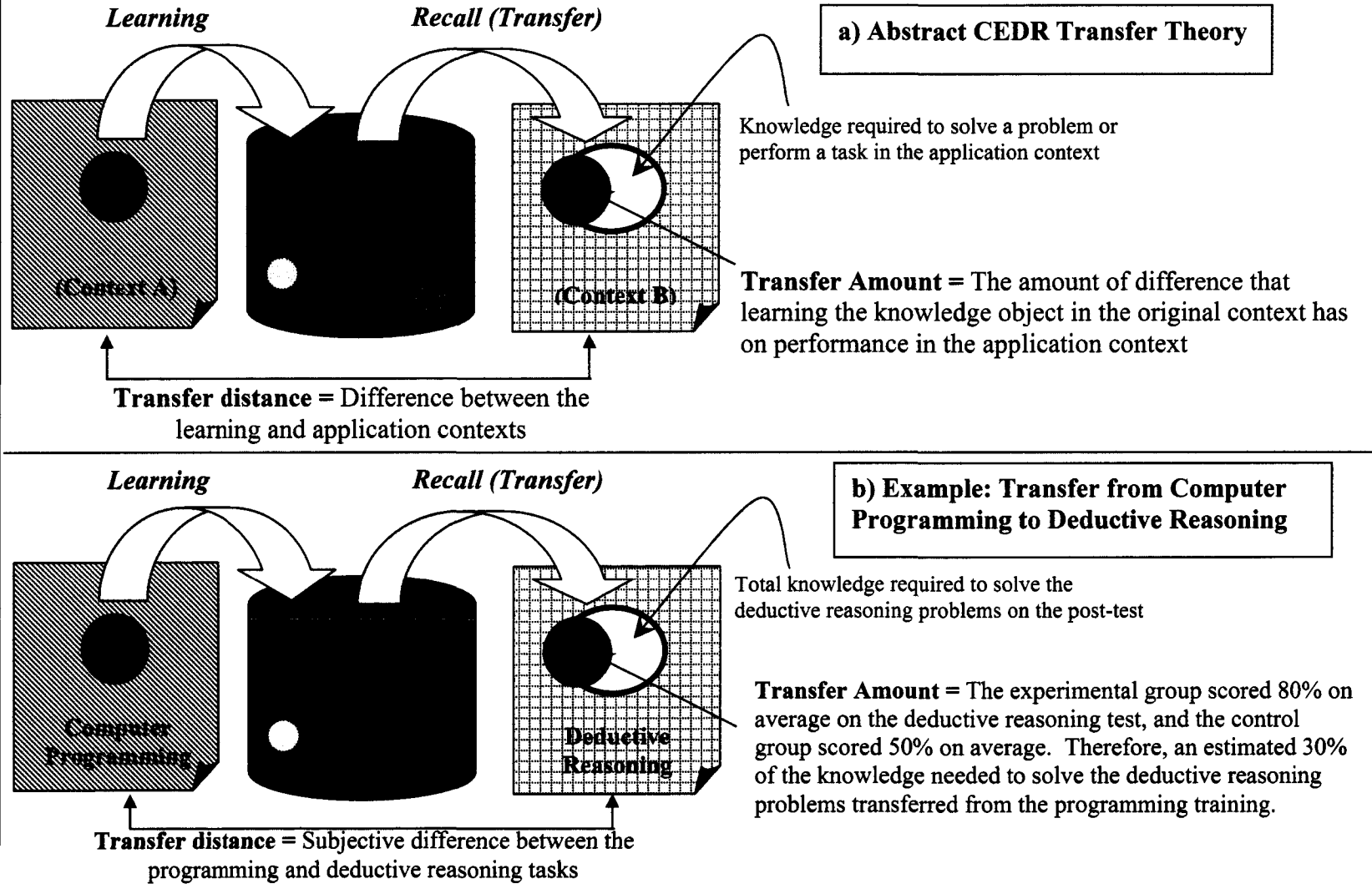
- Why do we see a lot of near transfer and not a lot of far?
- How can *A* transfer to *B* more or less than *B* transfers to *A*?

I revisit these questions in the discussion section.

A CEDR model of knowledge transfer

The CEDR neural hypothesis explicitly supports a model of knowledge acquisition, storage, and recall. Not much else about cognitive processes can be inferred from this mechanism beyond these basic operations, however. In the context of knowledge transfer, in particular, the CEDR hypothesis does not shed any light on the differential roles of context (task domain) vs. content (knowledge objects) in cognitive processes, and it provides no explicit clues about the nature of physical mechanisms that might be involved in transfer. It is nevertheless (or perhaps I should say "it is therefore") quite straightforward to operationalize the main constructs of classical transfer theory (distance and amount) in terms compatible with the CEDR mechanism (Figure 4.6).

Figure 4.6: Knowledge transfer in a CEDR system: a) Theoretical framework; **b)** Application to a specific example.



Since the CEDR hypothesis does not intrinsically support a theory of context or transfer mechanism, these elements can simply be defined functionally—that is, in terms of the observed behavior itself—as they classically have been. For example, in an experiment designed to assess the transfer from training in computer programming to performance on specific deductive reasoning tasks, a researcher would subjectively evaluate how different the programming task is from the application task as a measure of transfer distance (Figure 4.6b). In this case, deductive reasoning tasks specified in terms of the programming language syntax would be relatively near transfer compared to deductive reasoning tasks couched in the context of a murder mystery narrative (e.g., where a series of clues allow an investigator to deduce who committed the crime, why, how, where, etc.).

Similarly, the amount of transfer would be assessed based on behavioral measurements. For example, if an experimental group is trained in computer programming and a control group is not, and the experimental group correctly solves 80% of the deductive reasoning tasks on a post-test while the control group only scores 50% on average, then the researchers might reason that 30% of the knowledge necessary to solve the deductive reasoning tasks transferred from the programming training on average. It is very straightforward to operationalize the behaviorally grounded constructs of classical transfer theory in terms of the CEDR framework because the latter neither informs nor constrains the former. In this sense, the theory of knowledge representation based on the hypothetical CEDR neural mechanism and the behaviorally grounded theory of knowledge transfer can be developed independently. Without additional relevant neural constraints, in fact, it seems like this option is the only one available.

A CNDR Model of Knowledge Transfer

The CNDR neural hypothesis supports a very different view of the same set of transfer phenomena. First, there is an obvious role for context in the CNDR system. If the system has many little machines connected in "assembly lines" (networks), then context would be necessary to switch on the set of machines relevant to the current task and switch off the rest (a kind of simultaneous priming effect). Roughly speaking, in this model it is plausible that contextual inputs configure the network of machines for the task and the task content itself provides the inputs to the resulting assembly line.

Second, the CNDR hypothesis intrinsically supports two different knowledge transfer mechanisms (Figure 4.7). The first is a spontaneous generalization mechanism (Figure 4.7a). In the case of a mechanical widget-making machine, suppose that the widget producer shows the machine design engineer a blue prototype widget made of a particular kind of plastic. From that example, the engineer constructs a machine. Once the machine is built, however, it is not limited to making blue plastic widgets. If red or green plastic is put in, red or green widgets are produced, respectively. If a different kind of plastic is used, virtually the same widgets are produced (a form of assimilation). Depending on the machine design, it might even be able to handle quite different raw materials like rubber or plaster (but probably not solid metal). The point is that the machine was designed based on a much more limited set of examples than it is capable of handling in practice, which is a type of spontaneous transfer.

The spontaneous generalization feature exhibited by the machine would operate analogously in a CNDR information processing system (Figure 4.8). Recall the earlier data set involving shoe sizes and math achievement scores (Figure 4.2). In the CNDR scenario, a set of five data points was used to construct the linear equation summarizing

Figure 4.7: Two knowledge transfer mechanisms in a CNDR system: a) Spontaneous generalization; **b)** Machine re-use.

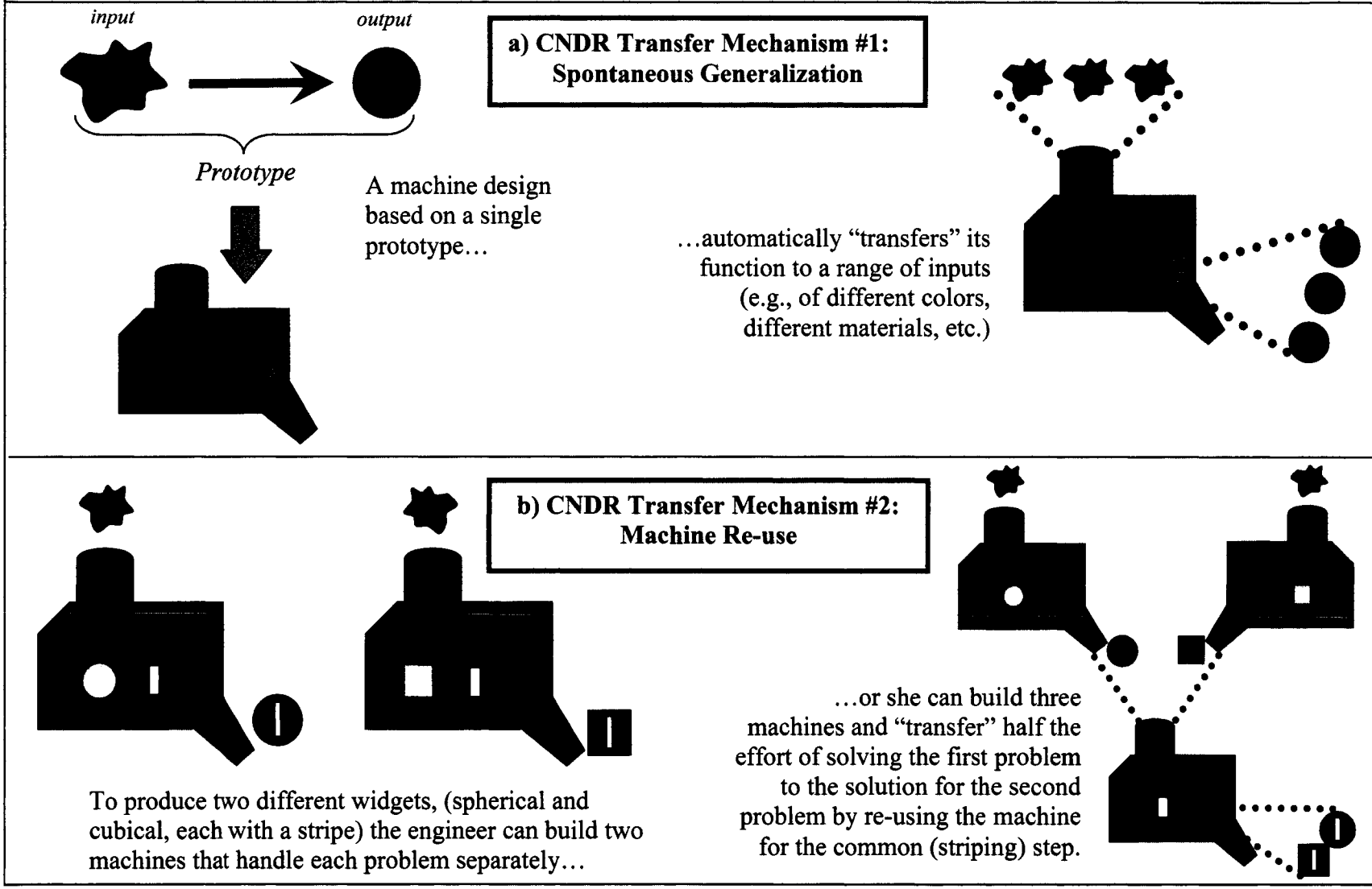
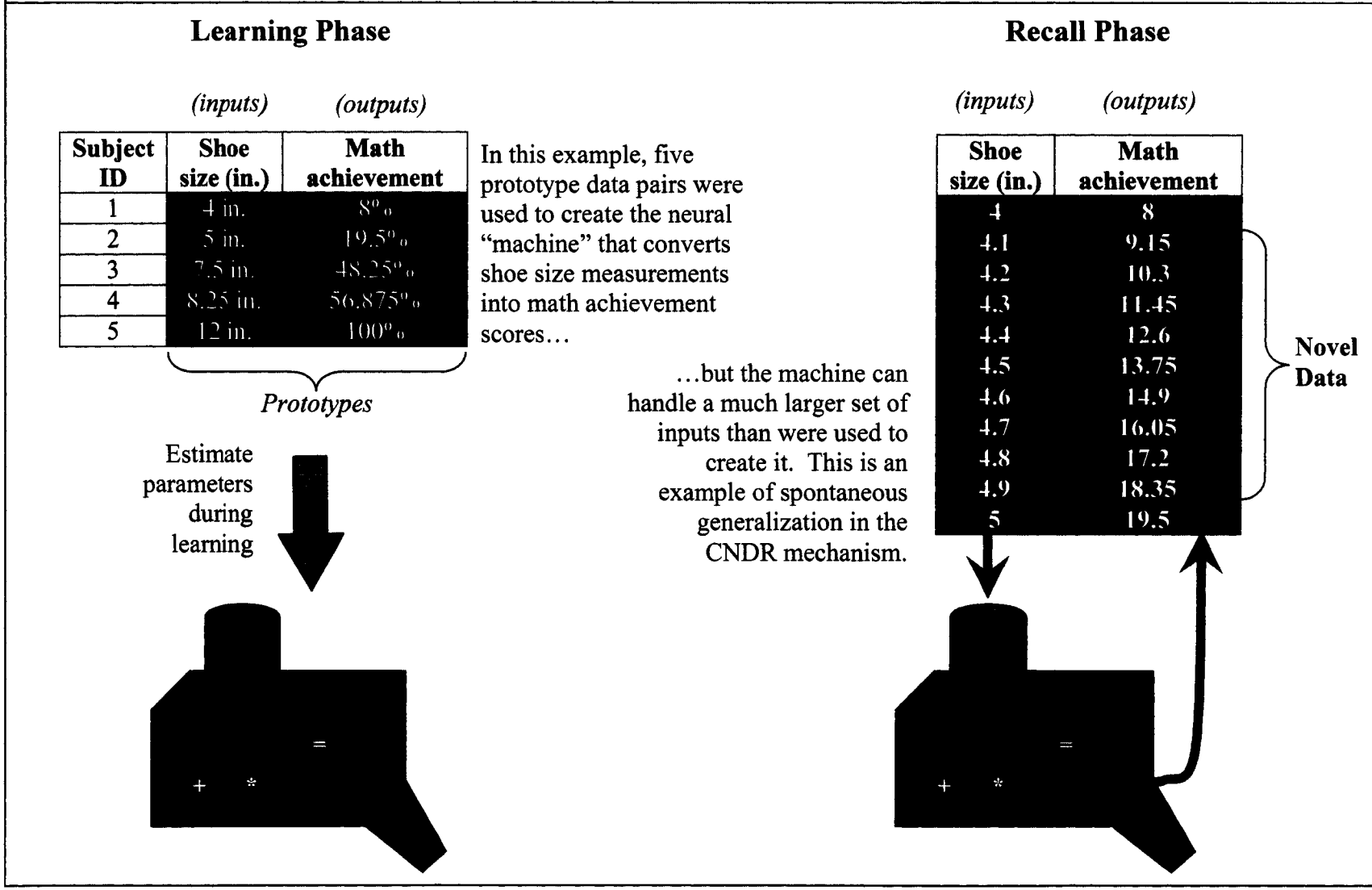


Figure 4.8: A concrete example of CNDR transfer mechanism #1 (spontaneous generalization)



the relationship between shoe size and math achievement. Unlike the CEDR system, the CNDR system does not store those five data points verbatim. It is consequently not limited to recalling those five "facts." Given any shoe size, the CNDR representation will automatically return an estimate of the associated math achievement score, in the same way the widget machine can process any color plastic. In this sense the knowledge encoded in a CNDR system is not distinguishable from the spontaneous generalization transfer mechanism. That is, the potential to generate the needed information embodied in the neural circuitry is simultaneously the actual knowledge that is encoded there and also a transfer mechanism.

The second transfer mechanism intrinsic in the CNDR system is based on machine-sharing across two or more sets of tasks (Figure 4.7b). Each machine in a CNDR system handles a small part of a problem, and in general many machines would be involved in any reasonably complex task. Returning to the manufacturing analogy, imagine the manufacturer decides to expand his line of widgets. In addition to spherical ones, he wants to produce cube-shaped ones as well. Finally, he wants to add a white stripe to all the widgets. The machine design engineer could build two independent machines: one to make striped spherical widgets and a second one to make striped cubical widgets. Assume it takes one month to design each processing step for each machine (e.g., one month for the sphere-making step and another month for the striping step). Building two independent machines would take two months for the first machine and another two months for the second machine. There is zero transfer in this case between the two tasks. If she is clever, however, the engineer will opt for a more efficient and flexible design involving three machines: one to make blank spheres, one to

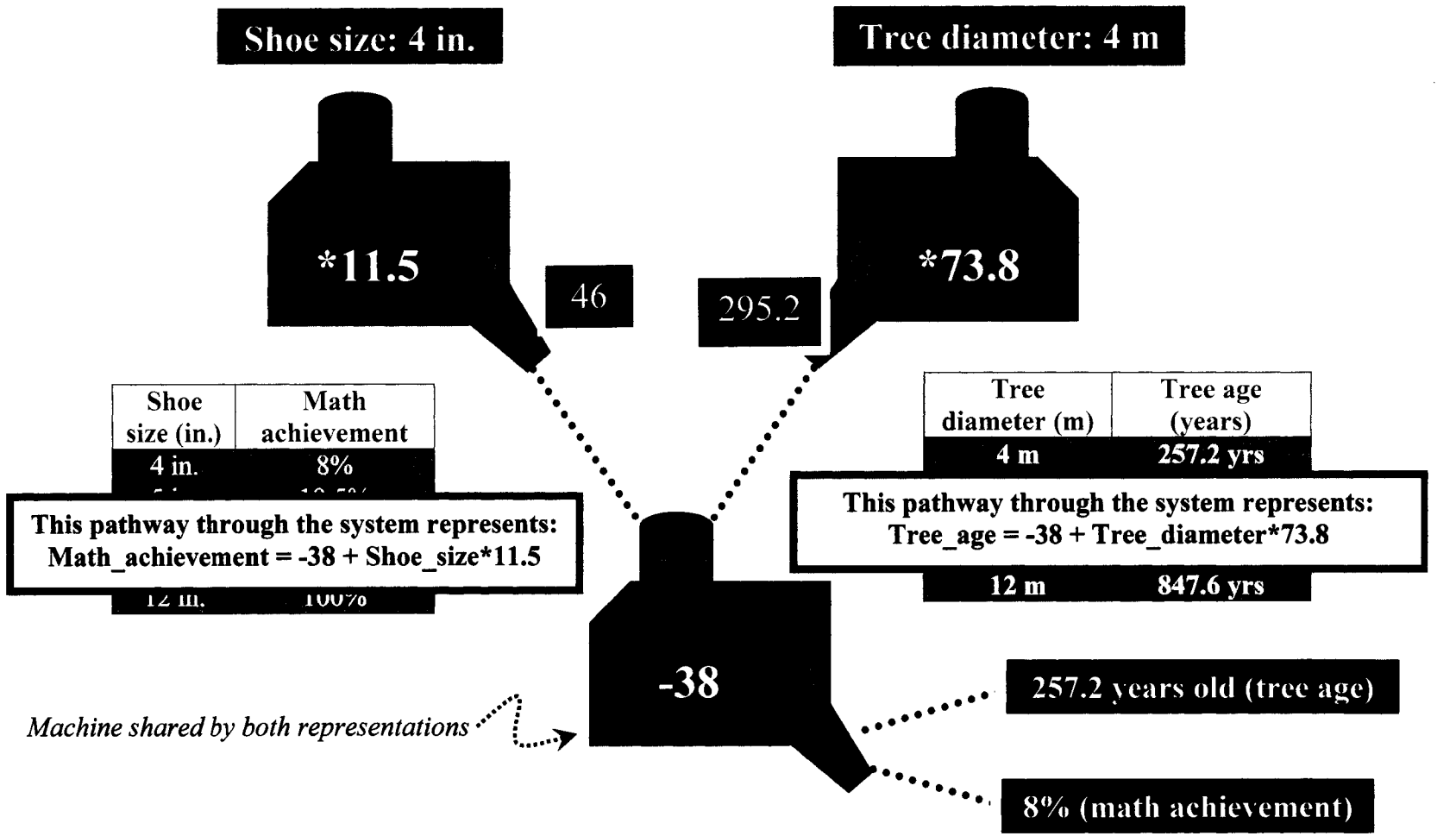
make blank cubes, and a third to paint stripes on whatever kind of widget is sent through it. This design requires two months for the first process (spherical widgets plus stripes) and only one month for the second process (to create cubical widgets only—she can reuse the striping machine from the first process). In this case, fifty percent of the design in the first task is transferred to the design of the second task by virtue of the shared machine.

A concrete example of the second CNDR transfer mechanism is illustrated in Figure 4.9. Assume our system has "learned" the data set on shoe size and math achievement using two machines. The first machine multiplies the input by 11.5 and outputs the result. The second machine takes the output from the first machine, subtracts 38 from it, and outputs the result. The entire process (represented by the left pathway in Figure 4.9) represents the data set on shoe size and math achievement. Now imagine that we want to represent a second data set involving measurements of diameter (in meters) and age (in years) of several members of a species of giant redwood. A linear equation summarizing this data is:

$$Tree\_age = -38 + Tree\_diameter*73.8$$

We need a new machine that multiplies its input by 73.8 and outputs the result. We can then feed the output from that machine into the machine we already have from the first process that subtracts 38 from its input. This process (represented by the pathway on the right in Figure 4.9) represents the data relating tree diameter to age. We could have constructed two separate machines that handled the two data sets separately, but the solution shown in Figure 4.9 is more efficient—it saves 50% of the effort of solving the second problem by re-using part of an existing solution. This is a second transfer mechanism, distinct from the spontaneous generalization mechanism. Note,

Figure 4.9: A concrete example of CNDR transfer mechanism #2 (machine re-use). Imagine we want to represent two different data sets that can be summarized with linear equations of different slopes (11.5 and 73.8) but the same intercept (-38). Instead of creating two independent neural circuits, we can build two circuits that share the machine that calculates the intercept.



however, that this mechanism will often make use of the spontaneous generalization mechanism. For example, the typical inputs to the second machine coming from the left-hand pathway in Figure 4.9 are going to be much smaller than the typical inputs coming from the right-hand pathway. When the input "4" is fed into both pathways, for instance, the input to the second machine from the left is 46 while the input from the right is 295.2. Since the "subtract 38" machine was created using examples from the first data set, it is operating out of its design range in the context of the second data set, which means that this case of machine re-use embeds within it a case of spontaneous generalization (but at a deeper layer of processing than the input layer).

Comparing the CEDR and CNDR knowledge transfer models

The CEDR and CNDR neural hypotheses support very different models of knowledge acquisition, storage, and recall. The CEDR model supports a knowledge "container" model, organized around discrete "knowledge objects." The CNDR model, in contrast, supports a knowledge "machine" model, organized in terms of implicit relationships among knowledge elements embodied directly in the neural tissue that are used to generate knowledge on demand.

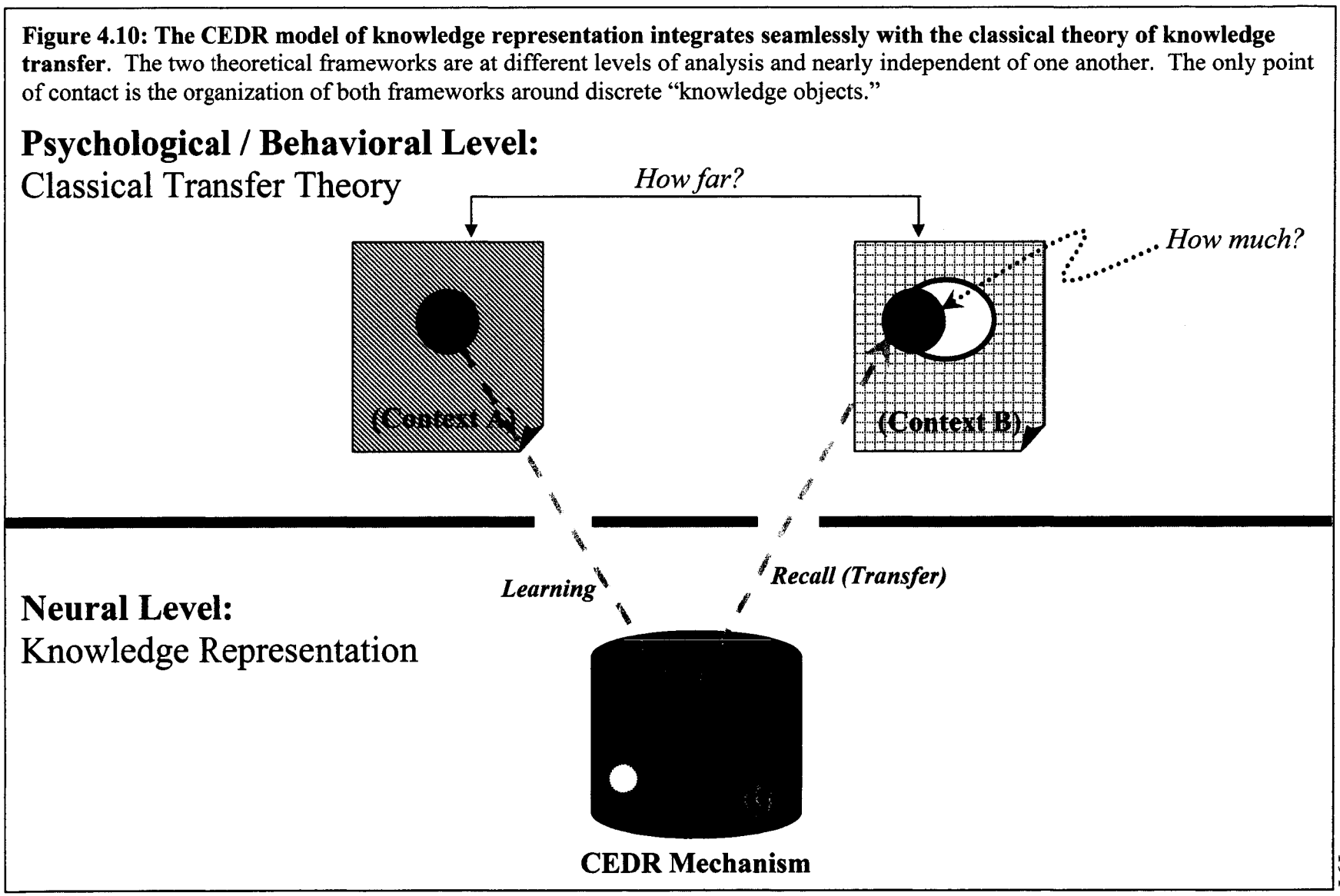
The CEDR neural hypothesis does not have any obvious direct implications extending beyond the knowledge representation model itself. Basically, the container model is a passive storage model in the sense that it simply records what is impressed upon it—it has no internal structure to speak of that might constrain knowledge organization or other cognitive processes (that is, beyond storage and recall). The passive character of the CEDR storage model combined with its organization in terms of

discrete knowledge objects makes it very easy to integrate with classical transfer theory (Figure 4.10).

The key elements of transfer theory are transfer amount (how much?) and transfer distance (how far?). Classically, transfer amount has been defined functionally in terms of behavioral measures, and transfer distance has been defined functionally in terms of the subjectively evaluated difference between the learning and application contexts (Salomon & Perkins, 1989). Because these key theoretical constructs are defined exclusively in terms of behavior, they are effectively insulated from phenomena at the neural level. Since the CEDR model does not "show through" in this case to the behavioral level and the classical transfer model does not "reach down" to the neural level, the theoretical primitives at these two levels can co-exist without conflict—the two levels are nearly independent of one another. The one point of overlap occurs because the classical transfer framework is, like the CEDR mechanism, organized largely in terms of knowledge objects. The very language of "distance" and "amount" of "transfer" suggests a metaphor wherein some object or part of an object is being transported a physical distance. This concurrence between the neural CEDR model and the behaviorally-grounded transfer theory in the fundamental unit of analysis (the knowledge object) is a factor that suggests the two frameworks are compatible even though they are otherwise independent.

The CNDR neural hypothesis, in contrast, provides a model of learning and recall that has wider implications for knowledge transfer. In particular, two transfer mechanisms emerge as side effects of the basic CNDR representational mechanism: spontaneous generalization and machine re-use. These mechanisms would operate at

Figure 4.10: The CEDR model of knowledge representation integrates seamlessly with the classical theory of knowledge transfer. The two theoretical frameworks are at different levels of analysis and nearly independent of one another. The only point of contact is the organization of both frameworks around discrete "knowledge objects."



every stage of processing from sensory inputs to motor outputs and including every layer in between. It stands to reason, therefore, that these neural mechanisms should be involved in the observable behaviors associated with knowledge transfer. In addition, the CNDR model provides a principled basis for distinguishing between context and content: task context configures networks of machines and task content activates the machines (that is, provides the raw inputs to them). These implications of the CNDR model represent a set of bottom-up theoretical primitives relevant to the behavioral phenomena associated with transfer. Classical transfer theory provides a different set of behavior-level theoretical primitives (distance, amount, and context) that deal with the same set of phenomena. These two sets of theoretical primitives are grounded in different kinds of data at different levels of analysis (neural organization and behavior, respectively), yet they potentially conflict where they provide different explanations for the same set of observable behaviors (Figure 4.11).

As an example of a specific conflict between classical transfer theory and the CNDR neural hypothesis, consider the classical continuum from "near" to "far" transfer in relation to the two CNDR transfer mechanisms (Figure 4.12). In at least some cases of near transfer (e.g., learning to drive a car transfers to driving a rental truck) the CNDR spontaneous generalization mechanism would clearly be the dominant one, operating at or near the sensory layer (that is, the raw inputs to the whole network). In other words, car driving skill transfers readily to truck driving problems because the two problem scenarios are very similar in the way they look and feel, supporting spontaneous generalization at the first layer of machines. In cases of far transfer, the dominant mechanism would be machine re-use (since part of what makes transfer "far" is the fact

Figure 4.11: The CNDR model of knowledge representation potentially conflicts with the classical theory of knowledge transfer

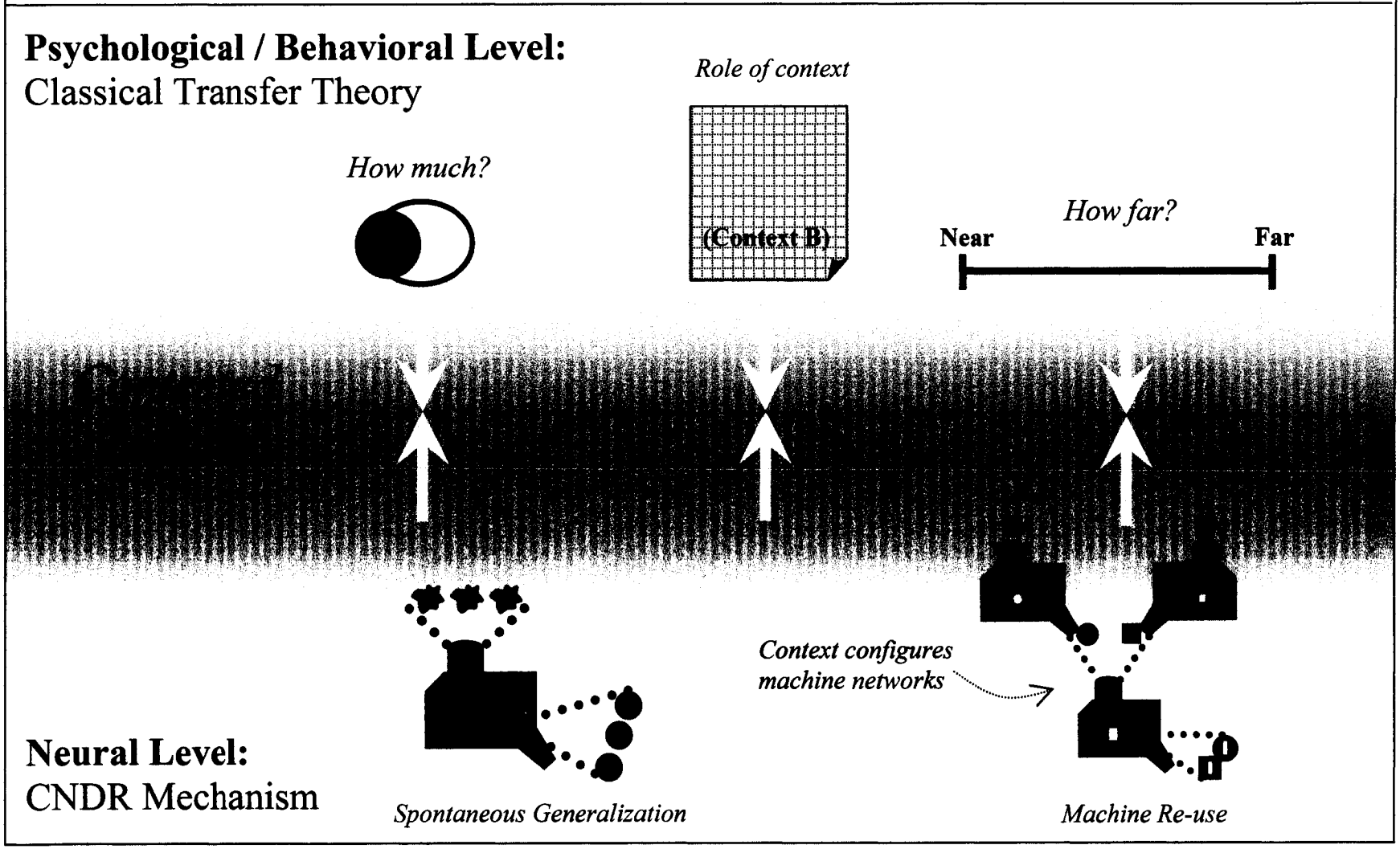
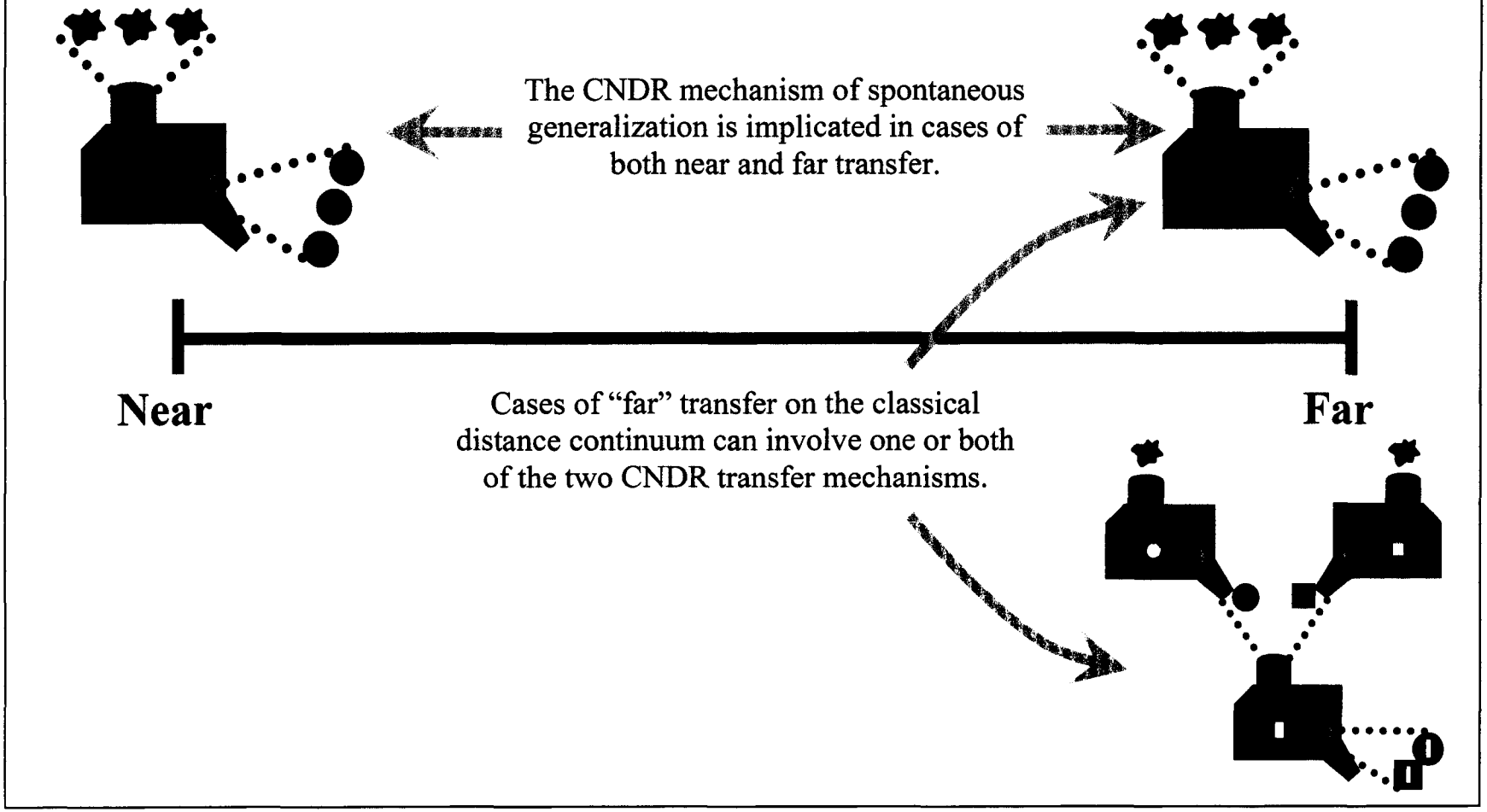


Figure 4.12: An example of a conflict between the bottom-up CNDR model of transfer and the top-down classical theory. The classical model posits a near-to-far continuum of transfer distance. The CNDR model gives rise to two transfer mechanisms: spontaneous generalization and machine re-use. These two sets of theoretical primitives represent competing explanations for a single set of behavioral phenomena, and they are not easily reconciled with one another.



that the two tasks are not perceptually similar so spontaneous generalization cannot be dominating at the sensory input layer). As I mentioned previously, however, the same spontaneous generalization mechanism does operate in cases of far transfer, although at deeper layers of processing (recall Figure 4.9).

This example highlights several points of incompatibility between the two theoretical frameworks. First, the classical "near" to "far" continuum is one-dimensional, whereas a measure of transfer distance based on the CNDR framework would be at least two-dimensional (since there are two different mechanisms involved). In other words, the classical continuum collapses across two qualitatively different neural mechanisms to create a one-dimensional behavioral yardstick for measuring transfer distance. There is, of course, nothing inherently wrong with making such a theoretical simplification in general. This particular case does seem to be problematic, however. For example, one of the longstanding open questions in transfer research is "Why do we see a lot of near transfer and not a lot of far?" (Salomon & Perkins, 1989). This question is particularly vexing to application-oriented researchers who want to design more effective educational materials and experiences by increasing the amount of far transfer that is induced by them.

Viewed from the CNDR perspective, this puzzle appears to be an artifact of the way the classical "near" to "far" continuum is constructed. That is, a one-dimensional behavioral yardstick implies a single underlying neural mechanism that operates uniformly in cases of near and far transfer alike. When transfer does not occur uniformly across a wide range of "distances," therefore, the empirical results conflict with the theory, giving rise to the puzzle. This puzzle basically dissolves when viewed through

the CNDR lens. In this view, "near" transfer is subserved primarily by the spontaneous generalization mechanism and "far" transfer is subserved primarily (but not exclusively) by the re-use of existing neural circuits. Transfer at intermediate distances would rely on various blends of the two mechanisms. As such, this framework does not lead to the expectation that "near" and "far" transfer should operate uniformly. The puzzle is thus seen to be an artifact of the way classical transfer theory is formulated (which could explain why decades of research have not been able to resolve it) that does not arise within the CNDR framework.

A second point of conflict between the classical and CNDR views stems from their fundamentally different units of analysis. The classical framework is organized around knowledge objects, and the distance continuum is based on this unit of analysis. Physical objects have the property of translational symmetry: carrying an object from point "A" to point "B" does not affect the object in a different way compared to carrying the object from point "B" to point "A." This symmetry property carries over metaphorically into the world of knowledge objects. For example, in the imaginary experiment described earlier in which transfer was measured from computer programming (context A) to deductive reasoning (context B), a knowledge object accounting for 30% of the deductive reasoning performance was transferred from the programming context. If the experiment were run backward, so that subjects received training in deductive reasoning (context B) and then were tested on algorithmic or computational problems (context A), one would expect from classical transfer theory that the same knowledge object from the first experiment would in this case be learned in the deductive reasoning context and transferred back to the programming context. This kind

of symmetry does not seem to hold in practice, however, leading to the longstanding open question, "How can *A* transfer to *B* more or less than *B* transfers to *A*?" (Salomon & Perkins, 1989).

Viewed from the CNDR perspective, this second puzzle is also seen to be an artifact of the way the classical transfer framework has been constructed (i.e., in terms of knowledge objects). The fundamental CNDR unit of analysis is not the knowledge object but the knowledge generating machine. A widget-making machine is specifically designed to operate in one direction: it might take molten plastic as its raw input, for example, and produce a spherical widget as its output. Such a machine cannot typically be run backward, however—if one were to put a spherical widget into either end of the machine one would not get molten plastic back out. Thus, the CNDR representational unit of analysis does not exhibit the symmetry property attributed to knowledge objects in the classical transfer framework. Since the CNDR transfer mechanisms emerge from the fundamental representational organization of this kind of system, there would be no reason *a priori* to expect transfer to be symmetrical, either.

For example, a CNDR network trained to play chess at an expert level (context A) might construct a network of machines capable of converting a static board configuration into an appropriate next move. The context in this case is well-defined (all chess games have a lot in common), and the content (e.g., the pieces, the board configuration, and the piece capabilities) only has a few dimensions of interest and is also well-defined. The chess-playing CNDR network would therefore have a few inputs (corresponding to the relevant aspects of the task) and the context ("chess game") would be able to configure

the network reliably for this particular task because there is little ambiguity in the relationship between context and content.

The domain of business (context B) has a different structure from the domain of chess. The context is more diffuse ("business situation" evokes many more possibilities than "chess game"), the content is more diffuse (there could be a large number of relevant problem features and they can change over time—for example, individual people have capabilities much more complex than the capabilities of chess pieces, and they are always learning new skills), and the relationship between context and content is ambiguous and variable (the task variables relevant to one problem instance might be different from the variables relevant in a different instance of the same kind of problem). The network of machines built to represent business strategy knowledge would therefore have many more inputs than the chess machine, and the activation of any particular machine configuration in response to a particular problem would tend to be less reliable, or at least much more nuanced.

Thinking in terms of transfer between the two contexts, the CNDR scenario is very different from a CEDR view based on knowledge objects. There are no decontextualized knowledge objects that can be passed symmetrically between two different contexts—there are only two very different networks of machines. If a business problem is fed into the chess machine (or, in classical terms, if the chess knowledge is transferred to the business domain), then there is a decent chance that a subset of the business inputs can be matched reasonably to the chess machine inputs (people are like chess pieces, different potential business sites have different strategic value just as different areas of the game board do, etc.). Basically, the more complicated business

problem is being "projected down" onto the smaller chess playing network by either compositing or selectively paring down the inputs. Feeding a chess problem into the business strategy network, in contrast, involves a process of "projecting up" from the smaller chess network to the larger and more diffuse business network. The input to the business network would tend to be severely underdetermined in this case. Even if there is some subset of the machine network that would be applicable to chess, there is a good chance the impoverished input set would not be able to activate it.

In other words, the statement "people are like chess pieces" (when applying the chess network to a business problem) is not symmetrical with "chess pieces are like people" (when applying the business network to a chess problem). In the first case ("projecting down"), thinking of people as chess pieces activates all the knowledge relevant to chess pieces for potential application to people, because chess pieces are less complex than people. In the second case ("projecting up"), thinking of chess pieces as people leads to the further question, "which features of chess pieces are like which features of people?" The fundamental asymmetries in the CNDR system (at the individual machine level and also at the level of two different knowledge networks) thus do not lead to the prediction that A should transfer to B the same as B transfers to A, and therefore the puzzle arising in classical transfer theory ("how can transfer be unequal in the two directions?") dissolves when viewed from this perspective.

At first blush, it might seem that the incompatibilities between the classical transfer and CNDR frameworks could be resolved by simply translating the theoretical constructs of one framework (e.g., "near" and "far" transfer) into the language of the other (e.g., spontaneous generalization and machine re-use). This strategy does not seem

viable in this case, however. For example, the two CNDR transfer mechanisms cannot simply be identified with near and far transfer, respectively (recall Figure 4.12). As I discussed, spontaneous generalization would be implicated in both near and far transfer, and far transfer would depend on both mechanisms, so there is no straightforward way to translate the classical distance continuum into CNDR terms or vice-versa. The differences between these two theoretical frameworks appear to stem not from superficial incompatibilities but from fundamentally incommensurable paradigms.

Conclusions

The starting point for this analysis was the single neuroscience finding that the brain uses two distinct mechanisms to store information: synaptic connections and dynamic patterns of neural activity. Based on that observation, I described two logically possible and mutually exclusive hypotheses (CEDR and CNDR) concerning the relationship between the synaptic and activity-based representations. The CEDR hypothesis (Coordinated, Equivalent, Distributed Representations) assumes the synaptic and activity-based representations are basically copies of one another, which leads to a "container" model of knowledge acquisition, storage, and retrieval. The CEDR mechanism implies a passive storage system that simply records what is imposed upon it, and therefore it has no intrinsic properties that necessarily "show through" to the levels of psychological or behavioral phenomena. Consequently, this bottom-up container model of knowledge storage can be integrated readily with top-down psychological theories grounded in behavioral data, such as the classical theory of knowledge transfer.

The CNDR hypothesis (Coordinated, Non-equivalent, Distributed Representations), in contrast, supports a "machine" model of knowledge representation.

In this model, the synaptic representations determine the structure of neural circuits that embody knowledge implicitly, while the activity-based representations hold the explicit knowledge associated with sensory inputs, motor outputs, recalled memories, intermediate products of processing, etc. Unlike the CEDR model, the CNDR neural organization has implications beyond information storage and retrieval. For example, two mechanisms of knowledge transfer can be identified as direct consequences of the CNDR representational system: spontaneous generalization and re-use of existing neural circuitry. These mechanisms presumably would be causally implicated in observable cases of knowledge transfer. As such, these CNDR transfer mechanisms constitute a concrete example of bottom-up (neurally-grounded) characteristics that do "show through" at the psychological and behavioral levels. Moreover, I argued that the theoretical primitives supported by the CNDR mechanism for explaining knowledge transfer are fundamentally incommensurable with the theoretical primitives of classical transfer theory grounded in behavioral data.

The research question being addressed in this paper is, "Do different assumptions about the brain support qualitatively different theories of particular psychological and/or behavioral phenomena? If so, then how?" Based on the results of the preceding analyses, the answer is clearly affirmative—the CEDR and CNDR neural hypotheses support qualitatively different theories of knowledge transfer at the behavioral level. The CEDR assumption does not constrain or inform a behavior-level theory of transfer, but it is compatible with the classical theory. The CNDR assumption, in contrast, leads to a model of transfer that directly conflicts with—and appears to be irreconcilable with—the classical theory.

The nature of the difference between the two neural hypotheses in terms of their higher-order implications for psychology and behavior is illuminating, if perhaps a bit surprising. One might have expected *a priori* to discover either no psychological implications on either side or two sets of implications that could be compared and contrasted. Instead, the key difference turns out to be that one hypothesis (CEDR) has no such implications while the other (CNDR) does. It would seem that if the brain is a CEDR system, then the assumption that brain and mind can be studied independently is perhaps warranted. If the brain is a CNDR system, however, then that assumption is contraindicated and psychologists should start thinking about drawing on neuroscience as a source of constraints for their psychological and behavioral theories.

If the CEDR system is compatible with the brain-mind independence assumption as well as a body of existing psychological and behavioral research and theory, then it might seem prudent to simply assume the human nervous system is based on that plan and carry on as we always have. The problem with that strategy is twofold. First, evolution has endowed us with a particular kind of neural system (be it CEDR, CNDR, or some other) and we have to work within the constraints of that underlying reality. While we are free to choose the kind of theory we want to build, we are not free to choose the kind of brain we actually have. Second, one finding of this analysis is that the two brain-level theories considered here do not appear to be interchangeable. The CEDR neural model is compatible with a behavioral-level theory that is incommensurable with the behavioral implications of the CNDR neural model. Therefore, they cannot both be true simultaneously, and it stands to reason that we need to build on the one that is true rather than on the one that is theoretically convenient.

Finally, it might seem as if one can avoid this problem altogether simply by refusing to make any assumptions about the brain at all in one's psychological theorizing. The difficulty with that strategy is that developing a psychological theory while assuming nothing about the brain is the same as assuming the mind can be described independently of neurological considerations. And that assumption is equivalent to the assumption that the brain has no intrinsic properties that "show through" to cognition and behavior. But that last assumption is tantamount to assuming the brain is a CEDR kind of system. The bottom line is that one cannot avoid relying on certain assumptions just by insisting one does not rely on them, any more than one can sidestep the behavior-level implications of brain organization by assuming they do not exist.

Chapter 5

There's More than One Way to Bridge a Gap: A Response to Bruer's "Bridge Too Far"

Introduction

Educational researchers and practitioners have long been optimistic about the possibility that neuroscience findings will inform educational theory and practice. In recent years, significant advances in neuroscience—accompanied by a stream of articles in newspapers, popular magazines, and professional journals touting the putative educational implications of these findings—have made neuroscience references a staple in the literature on educational theory, practice, and policy (Bruer, 1997, 1999b). Unfortunately, most of the educational claims based on these findings range in credibility from highly speculative to totally unfounded to downright nonsensical or even incomprehensible. Examples include the movement to develop curricula specifically tailored to the strengths and weaknesses of the "right-brain" vs. the "left brain," and more recently the popular but unfounded notion that parents can stimulate neural development that will boost children's ultimate mathematical abilities simply by exposing them to the music of Mozart from an early age (the so-called "Mozart effect").

In an effort to debunk some of the more widespread myths and redirect the general dialogue in this area into a more promising channel, Bruer (1997) wrote an influential paper entitled *Education and the Brain: A Bridge Too Far*, in which he argues that neuroscience cannot now—and possibly never will—inform education directly,

because the knowledge gap separating neuroscience from education is too large. Bruer[1] concludes that cognitive psychology is a more appropriate basis for a theory of education and instruction than neuroscience, that neuroscience findings can only inform education indirectly, and that the only feasible indirect route between the two is the one that begins with cognitive psychology as a theoretical point of departure and then bridges from this perspective to education on one side and neuroscience on the other (Figure 5.1).

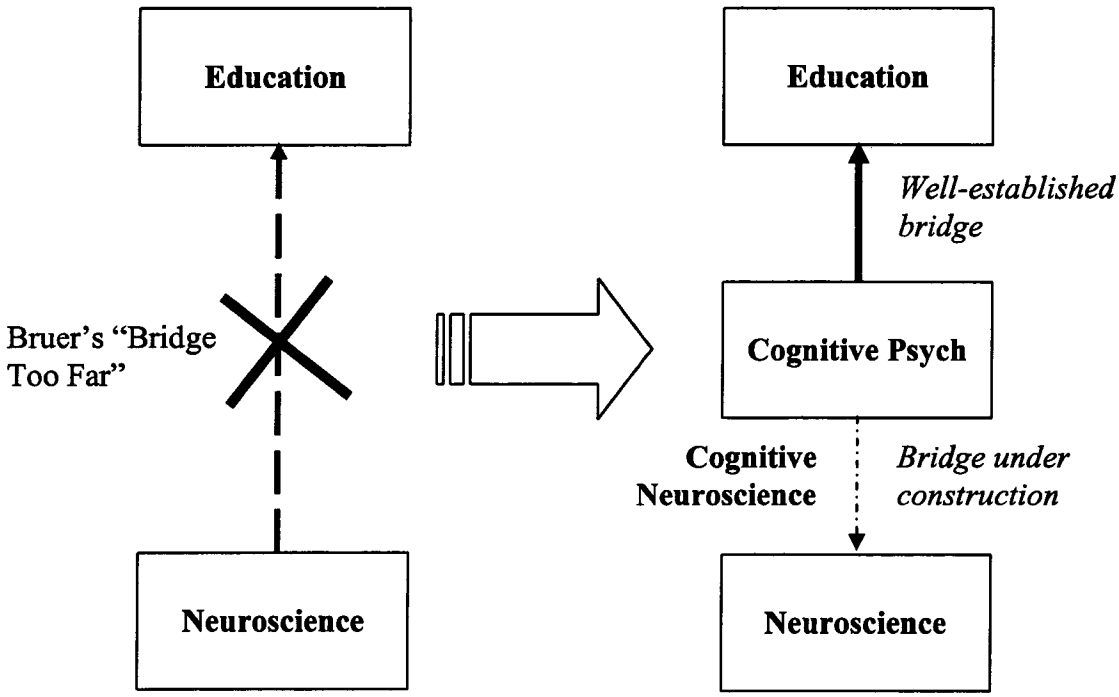
I agree with Bruer's critique of the existing neuroscience and education literature, and I hope that his cautionary message will reach the widest possible audience, encouraging educators, researchers, journalists, and the general public to adopt a more critical stance toward claims concerning the educational implications of particular neuroscience findings.

My view differs from his, however, on the current and future prospects for linking neuroscience to education. I would argue that Bruer's analysis is based on a conception of neuroscience that is too circumscribed, and as a result he overlooks some promising alternative links from neuroscience to behavior (and education). Specifically, in describing the bridge from cognitive psychology to neuroscience, Bruer seems to focus much of his attention on neuroimaging techniques within cognitive neuroscience to the exclusion of other neuroscience sub-disciplines. I will argue that computational neuroscience, in which computer models of the brain are employed to explore the brain-mind relationship, is a promising alternative approach for linking neuroscience to education.

---

[1] I should mention that my purpose is very different from Bruer's. Whereas his primary aim was to state a position on the neuroscience and education debate (Bruer, 2004, personal communication), mine is to clarify and evaluate some underlying theoretical issues behind the debate.

Figure 5.1: Schematic representation of Bruer's argument about neuroscience and education. His basic contention is that neuroscience cannot inform education directly, and therefore cognitive psychology is necessary as an intermediate level of analysis.



In this paper, I offer a fresh analysis of the relationships among neuroscience, cognitive neuroscience, cognitive psychology, and education that is organized in terms of levels of analysis instead of disciplinary boundaries. The result is, I believe, a different perspective from Bruer's on the "gap" separating neuroscience from education. In addition, I introduce computational neuroscience into the mix in an effort to illustrate how this relatively new framework relates to more established disciplines and approaches. I conclude by arguing that computational neuroscience is a promising avenue of research with the potential to inform educational research and practice in principled ways, even in the near future, and therefore deserves attention from educational researchers.

Building Bridges between Neuroscience and Education

The force of the argument depicted in Figure 5.1 derives in part from its efficiency in mapping out the relationships among the disciplines of neuroscience, cognitive neuroscience, cognitive psychology, and education while simultaneously suggesting how the various disciplines can be roughly identified with the three levels of analysis from brain ("neuroscience") to mind ("cognitive psychology") to behavior ("education").

The correlation between disciplines and levels of analysis is not perfect, however, and in my view this analysis masks important insights that are relevant to the neuroscience and education discussion. I therefore endeavor to construct a parallel

analysis in which I place the disciplines within an organizing framework based on three levels of analysis instead of the other way around[2].

Defining the Levels of Analysis

As a starting point for defining the three levels of analysis, consider the colloquial terms "brain," "mind," and "behavior." Behavior can be defined simply as any directly observable externalized action (including such experimentally elicited responses as linguistic utterances, button presses, eye movements, etc.).

The word "brain" is most closely associated with the pinkish organ situated inside the skull—the complex structure composed of smaller structures like cells, synapses, proteins, etc.

The mind can be defined in terms of the other two—roughly speaking, it is everything that comes "between" the physical organ of the brain and the externally observable behavior. That is, "mind" is an abstract category containing all the internal representations and processes not directly observable that enable behavior and that are ultimately instantiated physically in the brain.

A conflict arises at this point. The brain is most closely associated with the physical organ by that name, but the brain also has a functional aspect. The cells, synapses, and neurotransmitters generate entities like physical spike trains. These are measurable physical phenomena, and in that sense they should be considered part of the brain. However, these phenomena are information-carrying processes (or the products of

[2] The full analytic framework introduced in chapter 2—which defines levels of analysis in terms of inputs, outputs, representations, and transformation functions and also specifies mappings between levels—is most appropriate for analyzing well-defined and detailed theories and models such as the production system or ANN. In this paper I demonstrate how the basic levels-of-analysis framework from that chapter and the materialist approach of "tracking the information flow" it embodies can also be applied fruitfully even to compare and contrast disciplines—which tend to be more sprawling and ambiguous, encompassing many diverse detailed theories and models—by scaling back the level of detail from specific models to paradigmatic methods and tools characteristic of each discipline.

processes), not independently stable material structures like cells, and therefore they also participate in the mind category (we can refer to these physical brain processes collectively as the "brain-mind" to distinguish this description of the mind from alternative descriptions derived from other sources of data, such as behavioral observation—see discussion below).

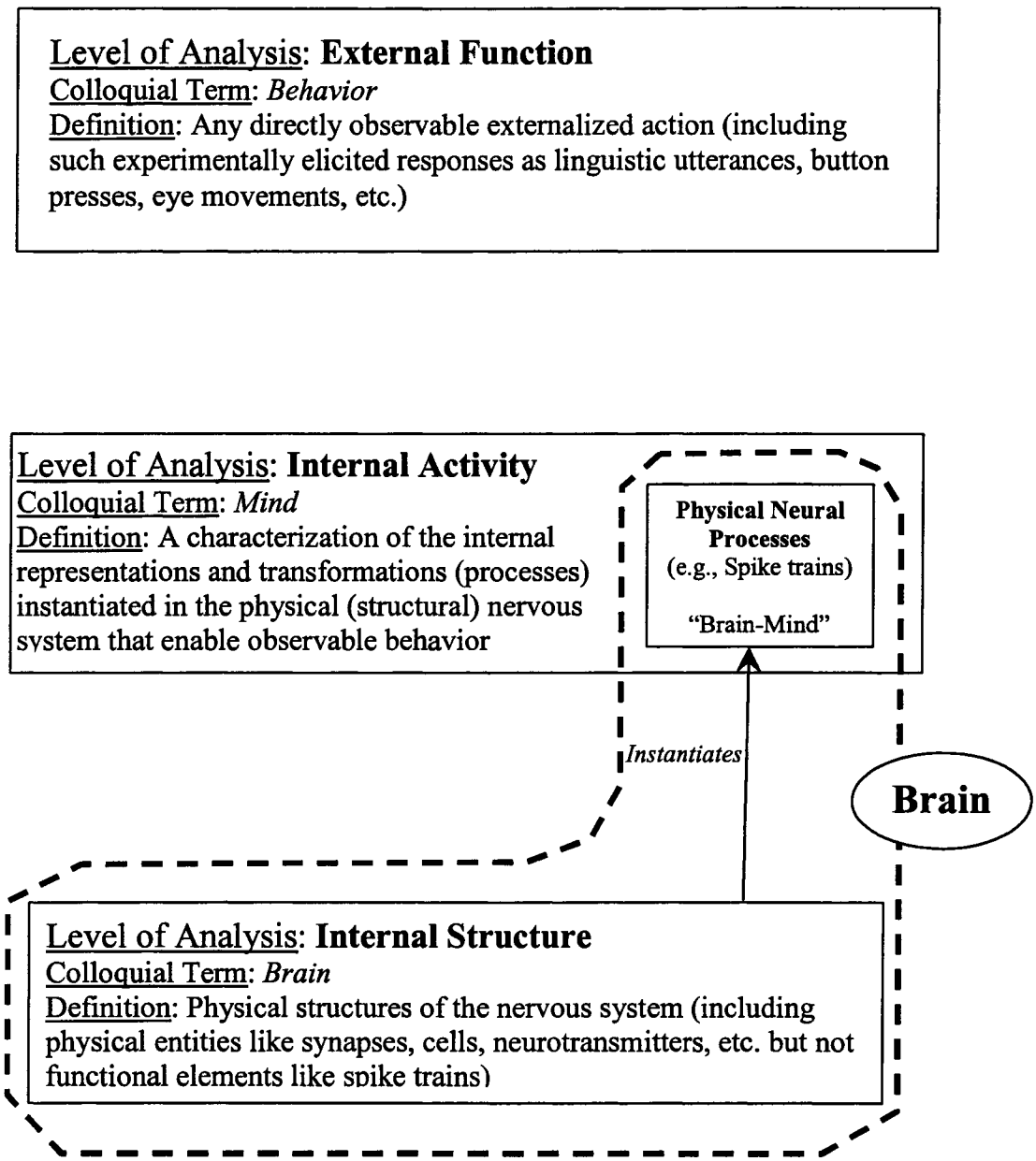
In order to understand the difference between structures (such as synapses) and activity patterns (such as spike trains), imagine scientists could flash-freeze a fully functional human brain without damaging it, simultaneously cutting off its energy supply and blocking its sensory inputs so all activity would cease completely. All the components of the brain that can be observed while the brain is in this frozen, inactive state (including synapses, cells, and neurotransmitters) are structures. All the phenomena that existed while the brain was active but disappeared at the moment it was frozen (including spike trains and the action potentials that constitute them) are activity patterns.

For my purposes, the distinction between physical entities (including synapses as well as spike trains) and functional categories (like "mind" and "behavior") is as important as the distinction between the levels of analysis, so I introduce the nomenclature specified in Figure 5.2 to preserve both. I continue to use the colloquial terms "brain," "mind," and "behavior" where this does not introduce any ambiguity into the discussion.

Reconstructing Bruer's Bridge

In this section, I examine the links between neuroscience, cognitive neuroscience, cognitive psychology, and education from the perspective of the levels of analysis defined in the previous section. First, I place each discipline from Figure 5.1 within my

Figure 5.2: Levels of analysis defined



organizing framework from Figure 5.2 (the result is illustrated in Figure 5.3), and then I discuss insights and implications following from this alternative analysis.

Cognitive Psychology: The Functional Architecture of the Mind

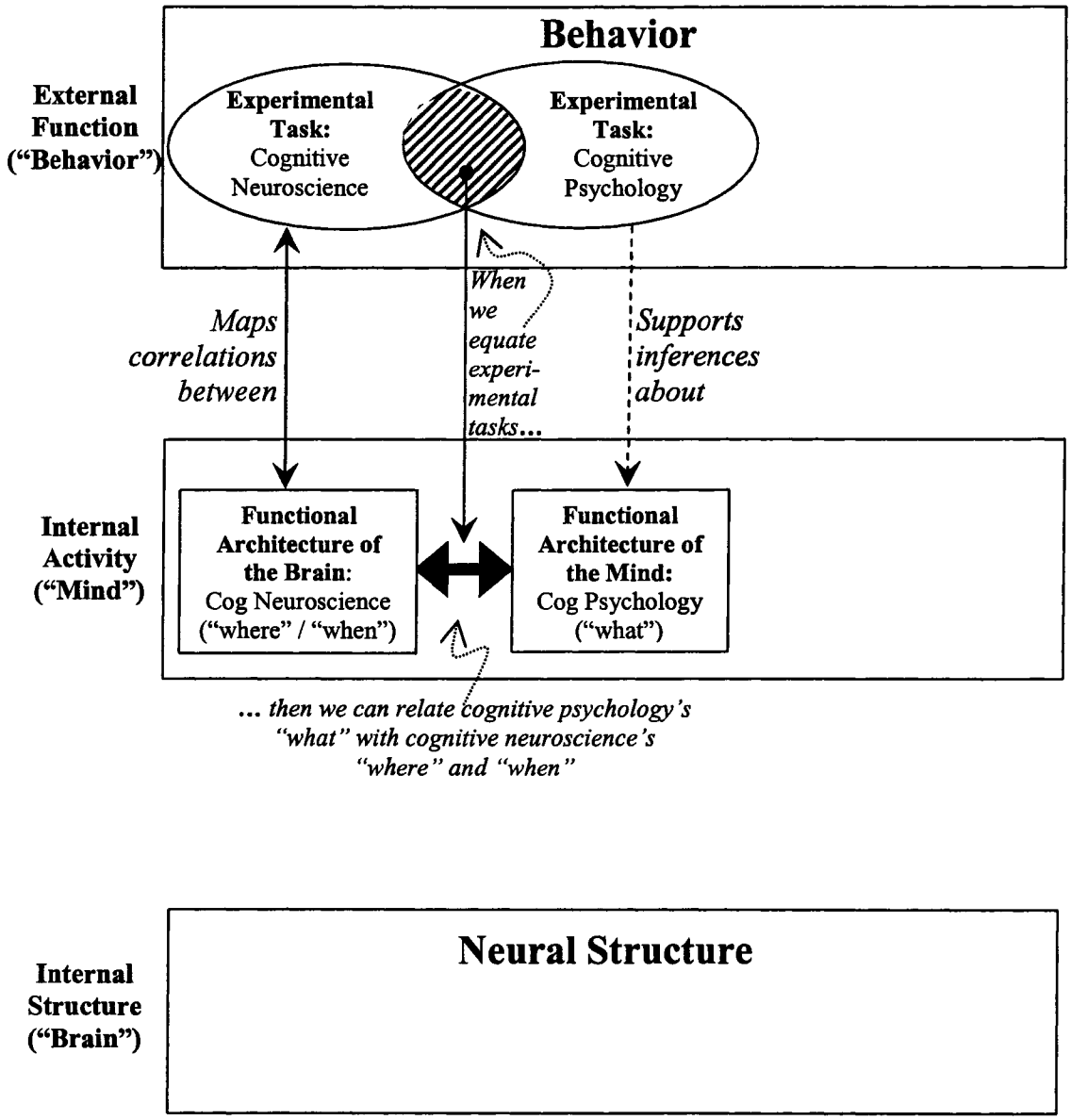
Cognitive psychology is "the study of mental activity as an information-processing problem" (Gazzaniga, Ivry, & Mangun, 2002, p. 97). The basic approach in cognitive psychology involves designing behavioral experiments to test hypotheses about the unobservable contents of mind: representations and transformation processes (Gazzaniga et al., 2002). In Figure 5.3, I have therefore placed cognitive psychology at the behavioral level, since that is where its data come from, with a dashed arrow pointing down to the level of internal function, indicating that from these behavioral data cognitive psychologists make inferences about the functional architecture of the mind—that is, about *what* abstract representations and transformations the mind contains, and *what* effect the transformations have on the representations, without regard for *how* or *where* those contents are physically realized in the brain.

Cognitive Neuroscience: The Functional Architecture of the Brain

Cognitive neuroscience is, generally speaking, the study of how the brain enables the mind (Gazzaniga et al., 2002). Although it encompasses a variety of experimental methodologies, this discipline is probably most closely associated with brain imaging techniques including fMRI, PET, MEG, and EEG. Indeed, from his examples this seems to be what Bruer (1997) primarily has in mind when he refers to cognitive neuroscience, so in the present discussion I restrict my comments to those methods.

The basic behavioral paradigm in brain imaging experiments is similar to the paradigm used in cognitive psychology. That is, in both cases a subject performs a

Figure 5.3: Schematic summary of my reanalysis of Bruer's argument (Figure 5.1), organized from the perspective of levels of analysis rather than disciplines



behavioral task. In addition to behavioral measures, however, in cognitive neuroscience brain activity is monitored using one or more of the technologies mentioned above. These activity patterns are widely assumed to represent the principal brain areas involved in the specific behavior under study (Gazzaniga et al., 2002). In other words, fMRI and PET scans provide data about "where" in the brain a particular behavior is processed and MEG and EEG provide data on its evolution over time ("when"). A major benefit of this approach is that it provides fixed points of reference (a brain map and time line) for comparing experimentally elicited behaviors with one another. On the one hand, if two behaviors activate roughly the same brain areas along a similar time course, then researchers infer that they involve some of the same neural processes. On the other hand, if two ostensibly similar behaviors activate different brain areas (either within a single group of subjects or across two different subpopulations) and/or evolve differently over time, then researchers conclude that the behaviors are supported by internally distinct processes, even though they appear similar externally.

Like cognitive psychological methods, neuroimaging techniques are grounded in behavioral data and therefore provide information about functional architecture. Unlike cognitive psychology, which typically must make inferences about the mind from behavioral data alone, these cognitive neuroscience techniques correlate behavior with brain activation patterns, and in this sense they give insight into the *functional* architecture of the *brain*. Neuroimaging technologies like functional MRI (as well as EEG/MEG and PET scans) "detect localized physiological *activity* within the brain, brain *function* ... rather than brain structure" (Churchland, 1995, p. 299, emphasis in the original). For these reasons, I have placed cognitive neuroscience (again, referring only

to neuroimaging techniques) at two levels in Figure 5.3 (external function and internal activity) with a solid double-headed arrow between them to indicate that these techniques correlate these two types of data.

Although they represent an important new source of data about the location and time course of processing associated with specific behaviors, note that these techniques still do not reveal *how* the brain actually implements these processes, either at the structural level (e.g., what role synapses and specific neurotransmitters play in the target behavior) or even at the level of brain function (e.g., how a specific pattern of neural firing encodes the target behavior).

Interlude: Reflections on the Bridge

Figure 5.1 summarizes Bruer's original discipline-based analysis, and Figure 5.3 summarizes my re-analysis of it, organized from the perspective of three non-overlapping levels of analysis spanning from neural structure to overt behavior. Two insights emerge at this point.

First, it is apparent that cognitive psychology and cognitive neuroscience are mutually complementary because they provide different kinds of information about the same general level of analysis (internal activity). If the experimental task is held constant (for example, a word-recognition task), then the behavioral data from a cognitive psychology experiment (information about mental contents) can be married to data from a brain imaging experiment (information about the sites where that information is processed, and the time course of that processing). This link is represented by the double-headed horizontal arrow in the middle layer of Figure 5.3. Generally speaking, cognitive psychology provides a rich database of hypotheses about the contents of the

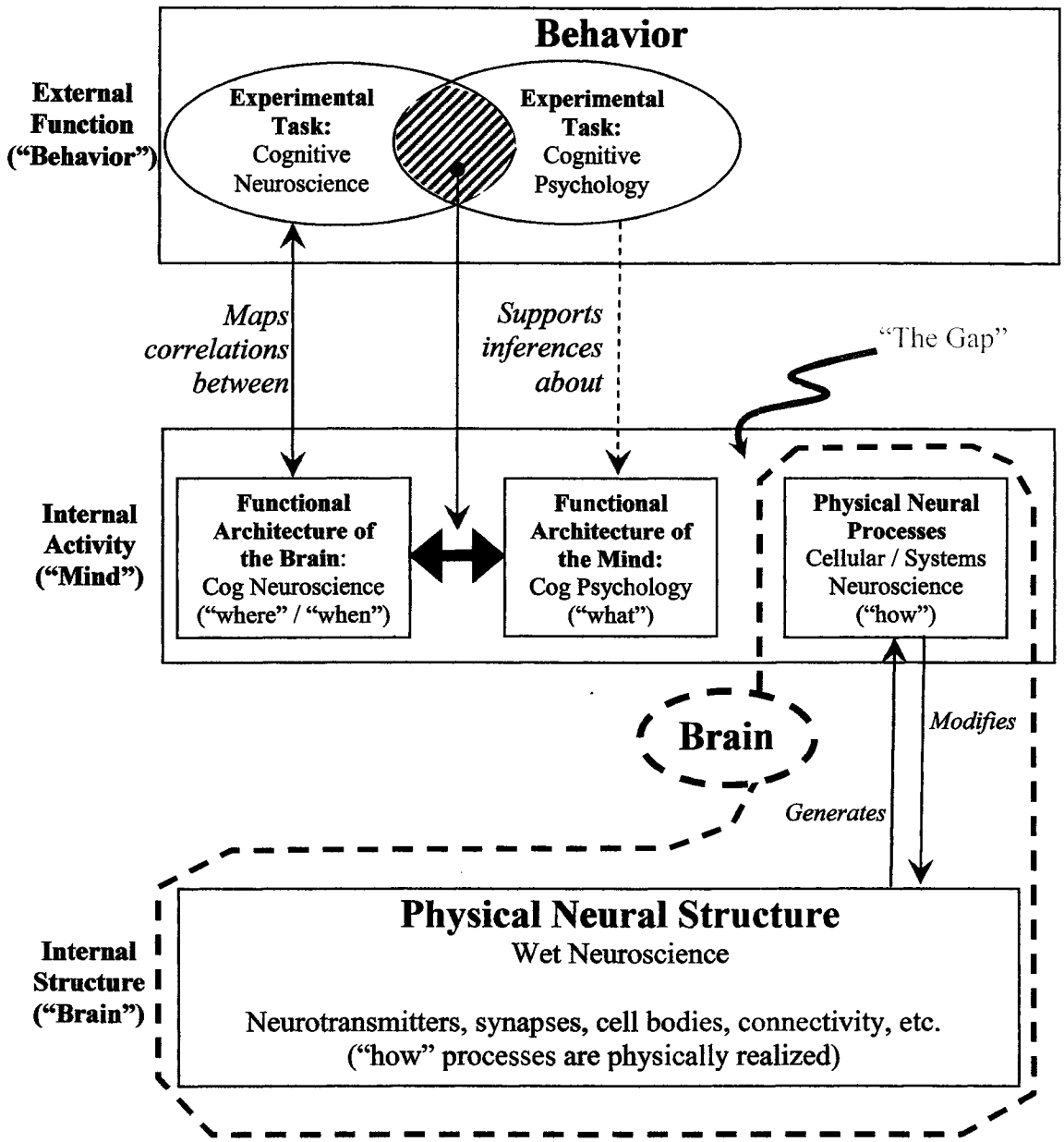
mind, while cognitive neuroscience provides powerful new methods for testing and refining those hypotheses, and each perspective is enriched by the exchange.

Second, the representation in Figure 5.3 suggests that the bridge in Figure 5.1 does not actually reach all the way down to the level of neural structure. The brain-imaging approaches are, in this view, better characterized as bridging from behavior to internal activity rather than from internal activity to internal structure (as Figure 5.1 might seem to imply). They do extend cognitive psychology—by grounding the inferences cognitive psychologists were already making about internal activity, but not by revealing *how* brain structures implement mental functions. Indeed, none of the disciplines represented in Figure 5.3 addresses the question of how the brain actually instantiates the processes that do the work of mental processing. Other branches of neuroscience are responsible for those kinds of questions.

"Wet" Neuroscience: Brain Structure and Brain Function

"Wet" neuroscience is the invasive study of the brain in a laboratory setting. Wet neuroscientists work at many levels of analysis, from the chemical structure of neurotransmitters in molecular neuroscience to the overall organization of the complete organ in systems neuroscience (Bear, Connors, & Paradiso, 1996). It is sometimes difficult to delineate where structure ends and process begins, but for present purposes it is sufficient to think of structures as the physical components of the brain—the properties and entities of the inert brain that can be studied reductively using tools like chemical assays, dyes, scalpels, and microscopes. These methods and data belong primarily at the "internal structure" level of analysis (Figure 5.4). In contrast, brain function can be studied *in vivo* using techniques like single-cell and cell array recording equipment, and

Figure 5.4: Relationships between major levels of analysis (external behavior, internal activity, and internal structure) and key disciplines within cognitive science



in vitro using various techniques that monitor physical, chemical, or electrical changes resulting when a stimulus is applied to a brain preparation (for instance, a brain slice). These methods and data belong primarily at the level of "internal activity" (Figure 5.4). An upward pointing arrow signifies that the brain structures generate the brain processes, and a downward arrow represents the fact that these processes can, in turn, modify the underlying structures (for example, as happens during learning). Together, these structures and processes constitute the physical brain (indicated in Figure 5.4 by the dashed outline labeled "Brain"). Note that the physical brain spans the levels of internal structure and internal activity in this framework.

Although wet neuroscience methods can be applied to a variety of questions, they are primarily distinguished by their ability to address implementation questions. For example, they address *how* proteins are used to build structures like ion channels that regulate the flow of charged particles into and out of neurons, and also *how* these particle flows contribute to the initiation and maintenance of spikes when a neuron fires.

Mind, the Gap

As Figure 5.4 illustrates, multiple disciplinary accounts of mental phenomena (that is, phenomena at the level of internal activity) are available. Cognitive psychology provides a language for talking about the contents of mind (representations and transformations), and brain imaging provides a basis for investigating the location and time course of processing. I discussed previously how these two accounts can be linked by holding constant the experimental task. As Figure 5.4 illustrates, these approaches move from the outside (external behavior) inward (toward internal activity) and their

theoretical formulations reflect this fact, being rooted in behaviorally-derived categories and concepts.

In addition, the brain itself embodies a third "language" for describing neural structures and mechanisms that actually do the work, some of which (for example, synapses, spike trains, long-term potentiation) have already been identified and are being characterized by researchers from various disciplines (see, for example, Alberts et al., 1994; Andersen & Koeppe, 1992; Bailey & Chen, 1983; Catterall, 1993; Hodgkin & Huxley, 1939). In contrast to the other two descriptions of internal activity discussed previously, this description moves from the inside (starting with brain structure) outward (toward internal activity and external behavior).

In my view, the gap alluded to by Bruer (1997) separating neuroscience from behavior and education (identified as "The Gap" in Figure 5.4) arises from the disconnection between these multiple descriptions of mind, some rooted in behavior and others rooted in neuroscience. Bridging this gap requires a way to translate the concepts of cognitive psychology into the language of neuroscience. For instance, "memory formation" is a cognitive psychological notion that might be translated into a neurological description involving gene expression at a set of synapses resulting from long-term potentiation arising in response to a novel stimulus. This would connect the descriptive "what" from cognitive psychology with the explanatory "how" of neural mechanisms. The feasibility of making this translation for many cognitive psychological constructs is controversial at present (Churchland, 1981; Churchland, 1988; Nunn, 1979; Vitzthum, 1995; see also chapter 4 for an explicit example of the kind of incompatibility that can arise between neurally- and behaviorally-grounded theories).

My examination of the disciplines in terms of their levels of analysis (Figure 5.4) suggests the following: although neuroimaging techniques from cognitive neuroscience do help ground and refine the theories of cognitive psychology, their combined reach still does not extend into the area of explanatory neural mechanisms (either functional or structural). Therefore, cognitive psychology might not be the ideal ground in which to anchor the bridges connecting neuroscience to education. Fortunately, the discipline of computational neuroscience offers a distinctly different kind of bridge that could complement these other bridges. In addition, educators and psychologists have already started traversing the computational neuroscience bridge.

Computational Neuroscience: An Alternate Route

The basic approach in computational neuroscience is to begin with data from molecular, cellular, and systems neuroscience and use them to specify a mathematical or computational model of neurons and neural networks in order to study how behavioral phenomena connect to molecular and cellular phenomena in the brain (Sejnowski, Koch, & Churchland, 1988). In most cases, these models are then simulated on computers to explore their structure and dynamics, in order to compare their behavior with observed behavior of biological nervous systems or to generate new hypotheses about the mechanisms underlying observed behavior. As a group, these models are often referred to as artificial neural networks (ANNs).

ANNs take many different forms, useful for studying a range of neural and cognitive phenomena at different levels of organization. For example, researchers interested in the details of synaptic and neural dynamics often use very complex, biologically realistic models that capture as many of the structural and functional details

of the individual neuron as possible. On the other end of the spectrum are highly simplified models, often used to study phenomena at higher levels of organization (e.g., language processing). This approach is followed by many researchers using ANNs to study higher cognitive functions and brain-mind connections. The connectionist model, a well-studied type of simplified model, is described below.

I wish to emphasize that I focus primarily on the connectionist model in the present discussion only because it represents the most accessible and convenient example from the larger domain of computational neuroscience. In particular, I do not mean to imply that the connectionist model is the only type of extant ANN, that it is necessarily the most informative type in terms of the brain-mind connection, or that it is the most biologically faithful (although I do believe it captures some critical characteristics of the biological system; see chapters 2 and 3 for a discussion of these issues).

The Connectionist Framework

The connectionist model is a specific kind of ANN. In order to explain how these models are derived from neurological data, it is convenient to focus on two key findings from neuroscience that inform the model design. These neuroscience facts are quite well established:

- Neurologically speaking, *learning* involves processes that modify the *structural synapses* via which neurons in the brain communicate with one another. Evidence suggests that this finding applies to many major brain areas and structures, and across all the major kinds of learning, including motor, associative, declarative, and episodic kinds of memory formation (Bear et al., 1996).

- In most nervous systems, usable *knowledge* of a stimulus is encoded in the dynamic and fleeting *functional pattern of activation* across a large number of neurons (Abbott & Sejnowski, 1999).

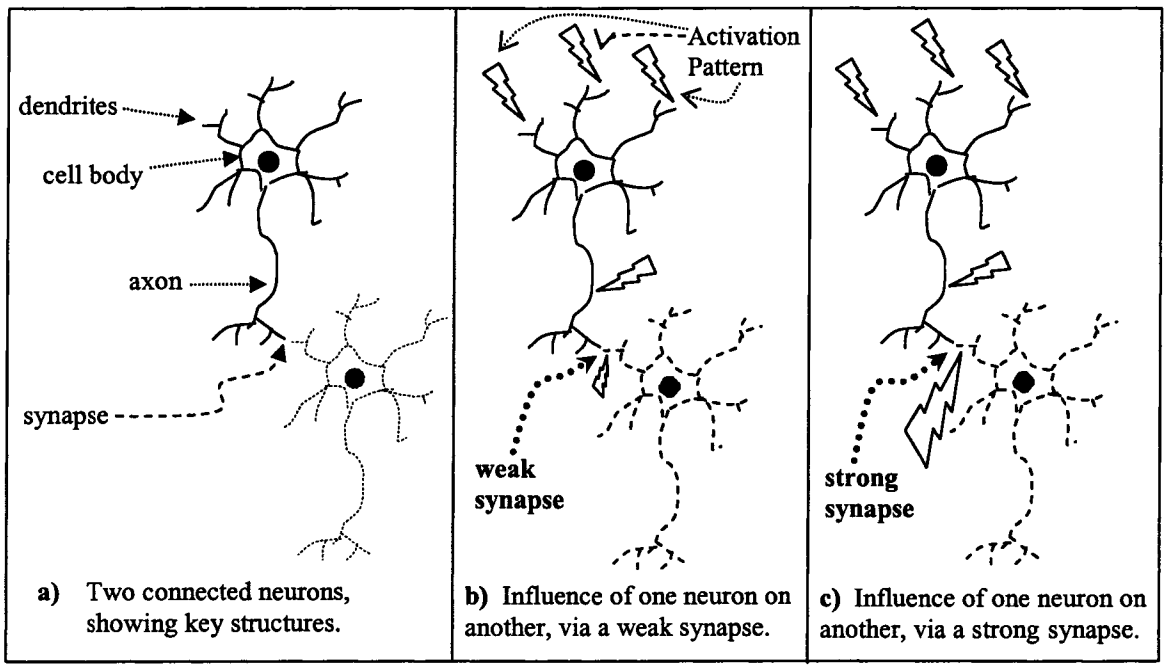
Figure 5.5 illustrates the relationship between structural synapses and functional patterns of activation in a neural network. Learning induces structural changes in a set of synapses (making them stronger or weaker). These structural changes affect the functional relationships between activation patterns at different points in the network, which determine what knowledge is stored in the network and how that information can be processed. Note that this thumbnail description embodies a simple explanation for how internal structures (that is, neural synapses) relate to internal functions (that is, neural activation patterns).

As I mentioned previously, *artificial* neural networks (ANNs) are mathematical models of *real* neural networks, like those that make up the human brain. Whereas the basic processing element in the brain is the *neuron*, the analogous element in a connectionist network is called a *node*. Figure 5.6 illustrates the correspondence between a biological neuron and a simulated connectionist node. Nodes are connected together to form networks, sometimes informed by data on neural connectivity patterns in the brain.

Although the nodes themselves are quite simple, the networks they form are surprisingly powerful. Indeed, computer scientists have shown that anything computable by the human brain is, in principle, also computable by some appropriately specified connectionist ANN (Chown, 2004; Hertz, Krogh, & Palmer, 1991). Because connectionist models are informed by neuroscience findings, they exhibit characteristics of the information processing that goes on in biological nervous systems (McLeod, Plunkett, & Rolls, 1998).

Of course, the connectionist model is intentionally very simplified compared to the biological nervous system. It is therefore not offered as a complete model of every

Figure 5.5: Neural structure and neural function. a) Key structures of a neuron include the dendrites, cell body, axon, and synapses. A neuron communicates by dumping neurotransmitters into the synaptic junction separating it from another neuron. These chemicals are detected by the dendrites of the neuron on the other side. b) The lightning bolts represent levels of activation at the neural inputs (dendrites) and output (axon). The size of a lightning bolt indicates the level of activation at a given site in the network, which is in turn controlled by the strength of the synapses through which it passes in traveling from one neuron to the next. *Learning processes* change the *strengths of synapses*. *Thought and action*, on the other hand, depend upon the *patterns of activation* across many neurons in the network. Synapse strength affects the influence of one neuron on another one, shown in panel (b) by the small lightning bolt beyond the weak synapse compared to the large lightning bolt at the same site in panel (c) beyond the strong synapse.

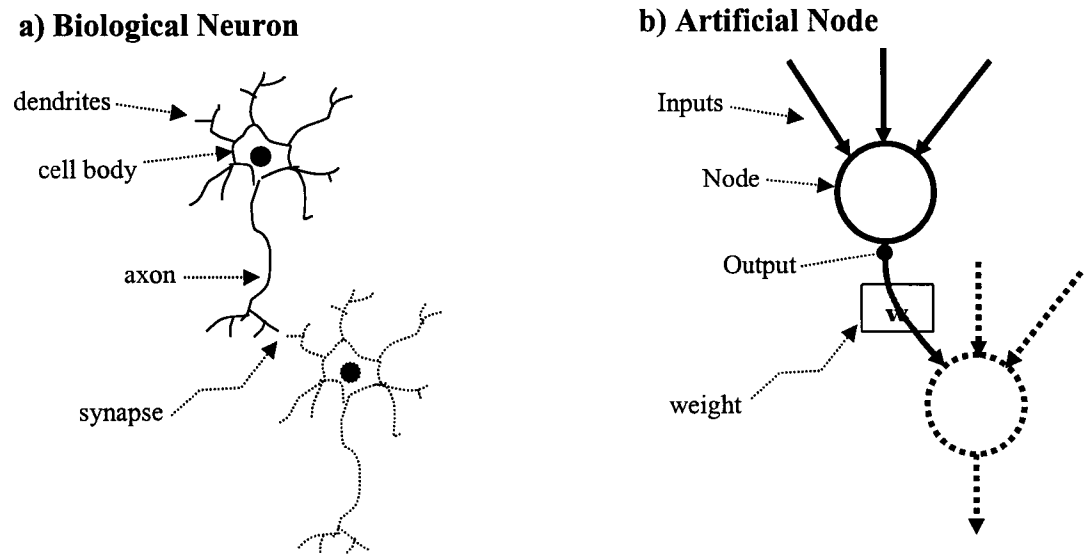


a) Two connected neurons, showing key structures.

b) Influence of one neuron on another, via a weak synapse.

c) Influence of one neuron on another, via a strong synapse.

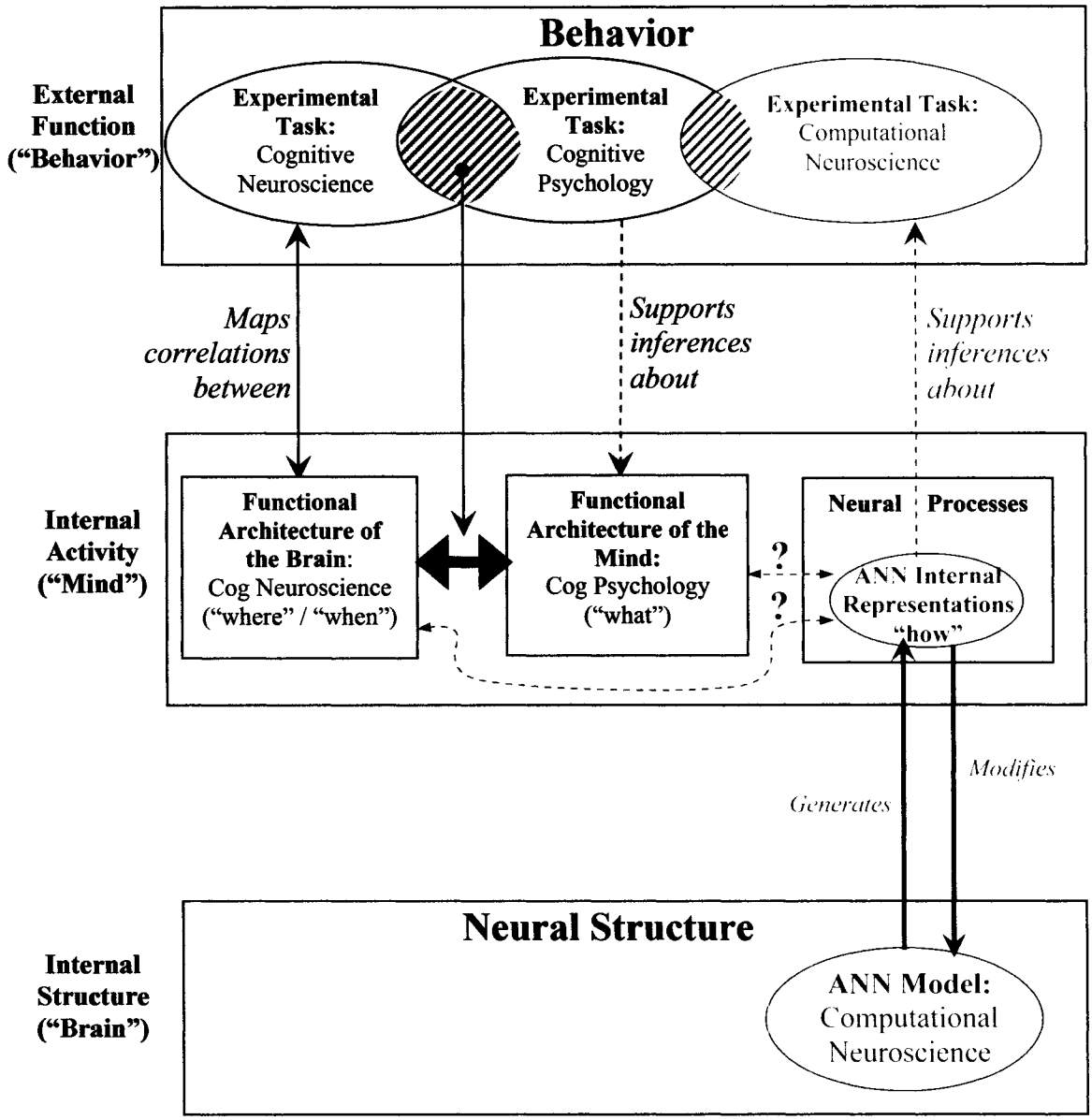
Figure 5.6: Key structures of a spinal motor neuron (left), and corresponding elements of an analogous node from an ANN (right). In the biological neuron, dendrites collect stimulation from other neurons. In nonlinear proportion to the total stimulation arriving on all the dendrites at one time, the neuron fires, sending activation down its axon and on to other neurons connected to it via synapses. The simulated node performs a similar operation. It sums its inputs, performs a computation on that sum, and sends the result on to other neurons. The efficacy of the biological synapse (which is modified by biological learning processes) is represented in the simulated model by the weight on the connection between two nodes (which is modified by simulated learning algorithms). The output from a neuron is modulated by the synaptic efficacy before being input to the next neuron in the chain, just as the output of a node is multiplied by the weight before being passed along to the next node in the chain.



process involved in human cognition. For example, connectionist networks often have no short-term or working memory, the sensory and motor systems are excluded, the modular architecture of the brain is generally not accounted for, and affective mechanisms are not explicitly included (though some parameters do crudely model some general influences of affect). Furthermore, as mentioned above, only a few aspects of the detailed synaptic or cellular behavior are captured in this framework. Nevertheless, many researchers think this is a good first approximation to the function of real neurons, (see, for example, O'Reilly & Munakata, 2000) and that the analytic simplicity of the model is a worthwhile tradeoff against the complexity that goes with strict neural realism.

To summarize, computational neuroscience is a branch of neuroscience that draws on wet neuroscience data to construct computer models of neural structures and processes, and the connectionist framework is just one example from the universe of computational neuroscience models. Researchers study the behavior of the resulting models via simulations to glean insight into the operation of biological nervous systems. This approach therefore links the level of internal neural structure to the level of internal neural activity, enabling researchers to investigate *how* the neural mechanisms implement cognitive processes (Figure 5.7). Moreover, model behavior is often used to make inferences about the internal functions generating externally observable human behavior (Abdi & Valentin, 1994; Addanki, 1984; Baker, Croot, McLeod, & Paul, 2001; Berg & Schade, 2000; Bollaert, 2000; Brady, 1995; Elman, 1989, 1993; Quinn & Johnson, 1997), represented in Figure 5.7 by the upward pointing dashed arrow connecting ANN internal representations to experimental task behavior.

Figure 5.7: Computational neuroscience represents a distinct bridge from internal neural structure to external behavior that bypasses cognitive psychology, although the two often draw on the same experimental tasks and paradigms



In addition, computational neuroscience models are considered a promising tool for integrating across many of the other neuroscience disciplines, including wet neuroscience (Kandel, Schwartz, & Jessell, 2000) and cognitive neuroscience (Gazzaniga et al., 2002; Kosslyn, Chabris, Marsolek, & Koenig, 1992). This point is represented in Figure 5.7 by the horizontal dashed lines linking the multiple descriptions at the level of "internal activity." As such, computational neuroscience could complement cognitive psychology and cognitive neuroscience by extending their combined reach to the level of internal structure. In addition, it is possible that ANNs could link neuroscience to behavior without making use of cognitive psychology at all, thereby offering an entirely distinct kind of bridge. Either way, computational neuroscience seems worthy of attention, as it offers a promising set of tools for crossing the neuroscience-education divide in the short term.

Conclusions

In his paper on neuroscience and education, Bruer (1997) identifies cognitive psychology as the discipline most central in applying principles from neuroscience to education:

> There are two shorter bridges, already in place, that indirectly link brain function with educational practice. There is a well-established bridge, now nearly 50 years old, between education and cognitive psychology. There is a second bridge, only around 10 years old, between cognitive psychology and neuroscience.... Cognitive psychology provides the only firm ground we have to anchor these bridges. It is the only way to go if we eventually want to move between education and the brain (p. 4).

In this paper, I have constructed a parallel analysis of the relationships between the disciplines of cognitive psychology, cognitive neuroscience, computational neuroscience, and wet neuroscience from the perspective of three distinct and non-

overlapping levels of analysis: external function ("behavior"), internal activity ("mind"), and internal structure ("brain"). This alternative analysis raises three challenges to Bruer's conclusions.

First, my analysis suggests that the "bridge between cognitive psychology and neuroscience" mentioned above does not actually extend into the realm of physical neuroscience to explain *how* mental processes are implemented in brain structures and functions. In my view, it is this gap separating the behavior-based descriptions of mental contents on the one hand from physically grounded explanations of brain function on the other that requires bridging, and both cognitive psychology and neuroimaging methods fall short of explaining how to accomplish this.

Second, a survey of extant philosophical positions on the brain-mind relationship (reductive materialism, functionalism, and eliminative materialism) reveals that the very idea that cognitive psychology can be reconciled with mechanistic neurological explanations of mind—even in principle—is controversial at the present time. In other words, it is not obvious that cognitive psychology represents an intermediate point on *any* path from neuroscience to education, let alone a necessary stop along *every* such path.

Third, an examination of computational neuroscience from the levels-of-analysis perspective supports the contention that this approach represents a distinct bridge from neuroscience to behavior that could ultimately bypass cognitive psychology altogether, challenging the view that the route through cognitive psychology is the *only* way to bridge from neuroscience to education. The emerging literature on ANNs and education suggests that educational researchers and practitioners are eager to traverse this bridge in their efforts to relate the behavior of ANNs to pedagogy, learning behavior, and

knowledge organization in people. Moreover, leading researchers in the domains of wet neuroscience and cognitive neuroscience point to computational neuroscience as the most likely way to integrate multiple theories of mind, to the extent that such an integration turns out to be possible.

This analysis suggests that computational neuroscience represents a viable bridge from physical neural mechanisms to behavior, and this route has several features to recommend it. For one thing, it is solidly anchored at each point: grounded at one end in well-defined neural structures and functions, and in observable external behavior at the other (in contrast to the one-sided grounding of cognitive psychology in behavior). In addition, computational neuroscience represents a single disciplinary framework that unifies the levels of organization from neural structures to behavior. Neural data must be translated into ANN model properties at one end and ANN activation patterns must be related to behavior at the other, but no paradigmatic translations are necessarily required in between. In Bruer's bridge, in contrast, multiple disciplinary boundaries must be crossed in making the trek, and each translation introduces new assumptions and layers of interpretation.

Perhaps most significantly for educators, the computational neuroscience bridge is already open to traffic. In the short term, ANNs are being used in instructional design and assessment, and also in some cases to provide the final span of bridge linking cognitive psychology and cognitive neuroscience on one hand to wet neuroscience on the other. These early efforts require support and validation from improved methods such as those discussed in the previous chapters. In the long term, computational neuroscience might provide a single unified theoretical framework supporting the principled translation

of neuroscience findings into educational practice. For all of these reasons, computational neuroscience and artificial neural networks warrant serious attention from members of the educational research community who are interested in applying principles of neuroscience to education in either the short or the long term.

Chapter 6

Conclusions: The Elephant in the Classroom (and What We Can Do About It)

The brain has become the elephant in the classroom. These days, virtually all educational stakeholders—from policy makers and administrators to teachers, parents, and even many students themselves—are aware that the brain is a factor in education, but no one really knows what to do about it yet. As a result, at the present time educational practitioners and policy makers have little choice but to ignore and work around it.

In this thesis, I have argued that excluding the brain from educational theory and design has material consequences. In chapter 4, for example, I described how two different assumptions about the brain lead to two incompatible psychological and behavioral theories of knowledge transfer. That analysis suggests that to the extent such theories inform educational designs and strategies, assumptions about the brain do matter in educational practice. More importantly, in that analysis I also pointed out why the issue cannot be sidestepped simply by remaining agnostic about the brain and refusing to make any assumptions about it at all. When one ignores the brain in that way, one implicitly assumes the brain tolerates being ignored. While some kinds of brains (e.g., CEDR) evidently can be pushed into a quiet corner and safely forgotten, other kinds of brains (e.g., CNDR) refuse to go quietly and continue to rampage about the room crashing into things and making their presence known. We are not free to choose our elephant—we have to work with the one we have been given. Until we know what kind of animal we are dealing with, we ignore it at our own peril.

Many people are already very interested in applying neuroscience to education, so it might seem as if I am preaching to the choir. I do not believe that is true, however. My argument is directed at scientists and scientifically-minded educational researchers. As Bruer (2002) points out, the vast majority of people interested in exploiting brain science for educational applications are not members of that community:

> Brain science may have caught the fancy of the media, policy advocates, and brain-based educators, but the educational research community appears to be considerably more reticent. A search of scientific databases (MedLine, SciSearch, and PsychInfo) reveals that published research on applications of brain science to general education is nonexistent. Furthermore, national commissions charged to review research for its relevance to educational practice see little of practical value emerging from brain science, even though these commissions include neuroscientists among their members (p. 1032).

I have reached basically the same conclusion as Bruer in this regard based on my own review of the literature. My appeal to the media representatives, policy advocates, and brain-based educators who continue to evangelize the putative educational implications of brain science prematurely would echo Bruer's in calling for far more restraint and rigor than has been exhibited to date.

My argument that brain assumptions (including the assumption that one can ignore the brain) have consequential implications for education is directed particularly toward those in the educational research community who are skeptical that neuroscience is relevant to education. By itself, that argument would merely tend to incite people without suggesting possibilities for action. In the bulk of this thesis, therefore, I have described a theoretical framework designed to support rigorous scientific research bridging from neuroscience to education. The framework is primarily meant to support actual research, but I also offer it as a proof-of-existence to another group of skeptics that

such research is even possible at the present time. The theoretical framework specifies the major steps involved in applying neuroscience to education (Figure 6.1), and includes a set of analytical tools, experimental methods, and concrete examples describing and demonstrating one implementation of the generation step and the translation step in that framework. The evaluation step is not discussed here; a large body of literature already documents methods for implementing it (e.g., through the design of controlled outcome studies of educational interventions). Together, these three steps are meant to establish a "pipeline" that takes basic neuroscience research as its raw input at one end and transforms it into validated educational applications that are output at the other end.

In the first step of the research process depicted in Figure 6.1, neuroscience facts are used to identify and characterize causal relationships between brain mechanisms and observable behavior (Figure 6.2). This is done by embedding the neural observation in a computational model (such as an ANN), identifying model behaviors that follow from the modeled mechanism, using these behaviors to generate predictions about human behavior, and then conducting experiments to test the behavioral predictions against human data. The experimental results are used as evidence for or against the hypothesized brain-behavior relationship in people. This method is in essence a framework for conducting basic research on brain-behavior links that is embedded within the overarching applied educational neuroscience research framework. I argued that this elaborated generation (and associated validation) step is necessary because the causal brain-behavior relationship is a more appropriate basis for informing educational interventions than is the raw neuroscience finding.

Figure 6.1: A minimal (three step) framework for conducting rigorous applied research in educational neuroscience

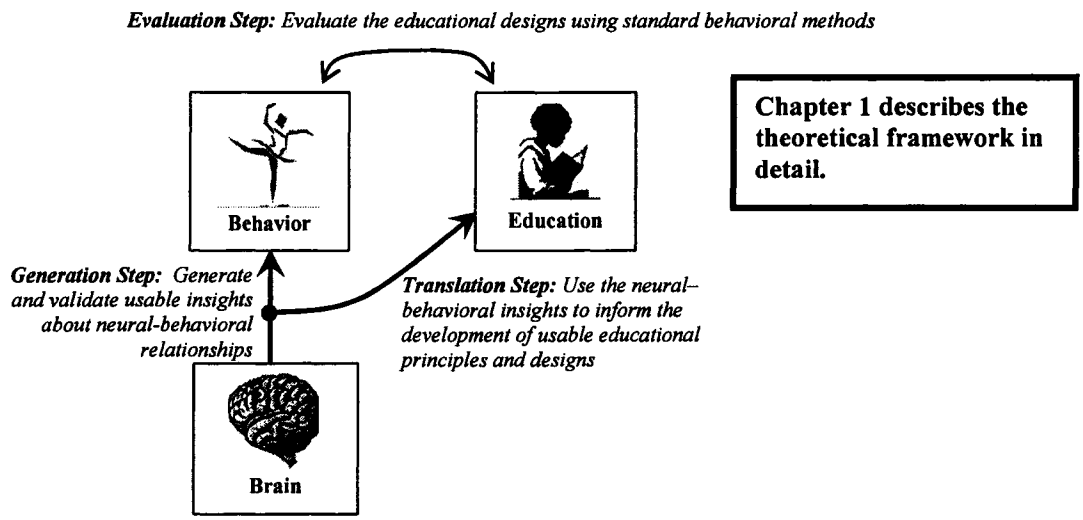
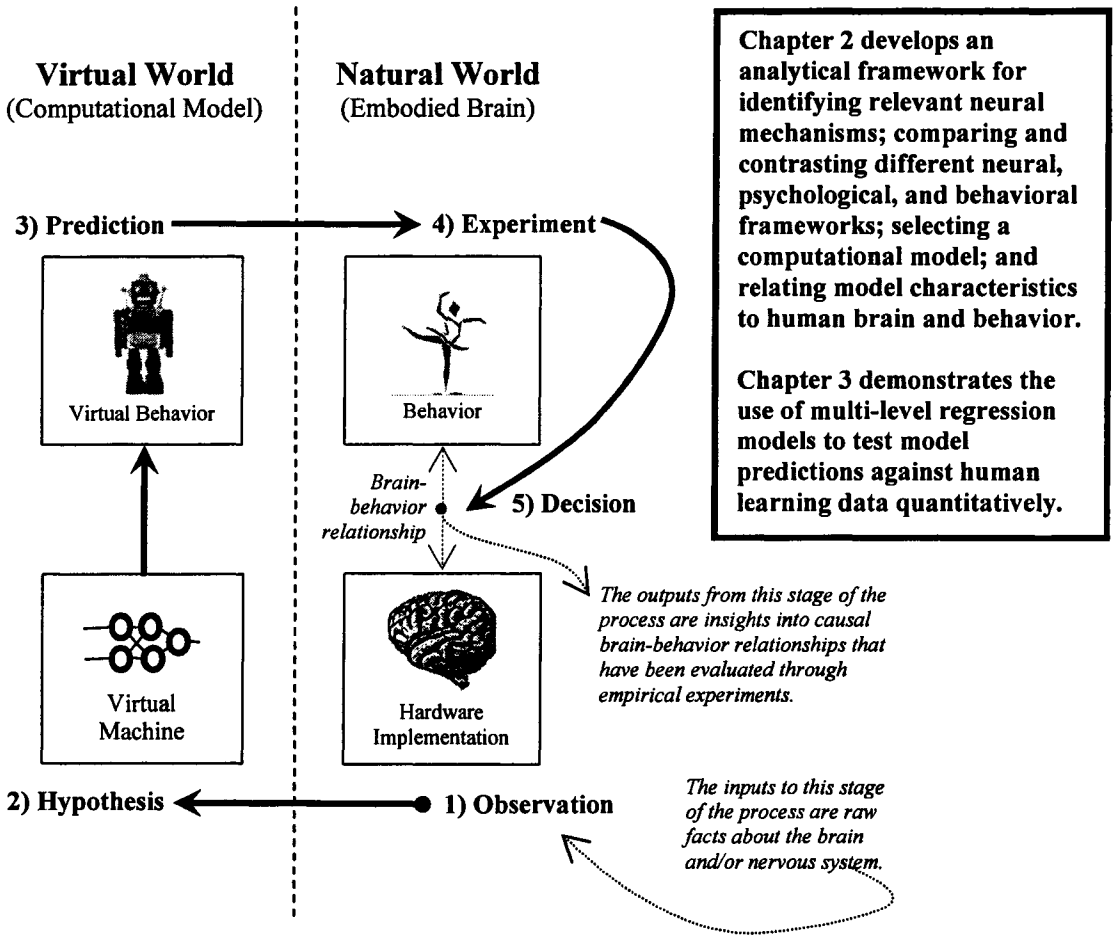


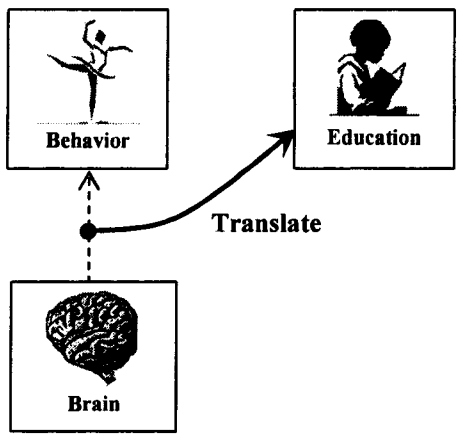
Figure 6.2: Generation step: Neural observations are used to generate candidate brain-behavior relationships.



In my case study, for example, I started with the observation that the brain employs two distinct types of distributed representations (synaptic and activity-based) that could in principle be coordinated in at least two distinct ways (producing the CEDR and CNDR mechanisms, respectively). I used an ANN to generate two sets of behavioral predictions (one about item difficulty and the other about changes in perception of item similarity associated with learning the dichotomous categories) following from one of them (CNDR). As a first step toward validating the causal brain-behavior CNDR mechanism, I tested the behavioral model predictions experimentally against human learning data using multi-level regression models. The experimental findings were consistent with both sets of model predictions, which I interpret as evidence supporting the CNDR hypothesis.

In the second step of the process (Figure 6.3), I translated behaviors identified in the first step into potentially useful educational insights. Specifically, I used analytical methods to link the hypothesized mechanisms (CEDR and CNDR) to patterns of behavior associated with knowledge transfer to argue that assumptions about the brain have consequences for theories of educationally relevant behavior. I also described a set of theoretical primitives derived from the CNDR neural mechanism that support a different view of knowledge transfer from the classical psychological theory. This alternative framework is organized around a completely different unit of analysis (the knowledge machine) than that used in the classical theory (the knowledge object). If elaborated, it could provide novel (or at least significantly refined) educational theories and design principles.

Figure 6.3: Translation step: The causal brain-behavior link identified in the first (generation) step is translated into usable educational theory and design principles.



Chapter 4 illustrates how a specific neural mechanism (the coordination between two sets of distributed representations) can be linked to educationally relevant behaviors (knowledge transfer).

In Chapter 5, the analytic framework from Chapter 2 is applied to Bruer's (1997) position on neuroscience and education to clarify the issues under consideration and establish a distinct position on the matter.

In the near term, the insights gleaned from the knowledge transfer analysis could be applied to design new kinds of transfer experiments. For example, the analysis suggests that one reason several fundamental unresolved puzzles about knowledge transfer persist could be because the theoretical primitives upon which the classical theory is based (transfer distance, transfer amount, and subjectively defined context) might be incompatible with the neural mechanisms actually involved in the phenomenon (e.g., spontaneous generalization and machine re-use). A knowledge transfer theory based on the CNDR theoretical primitives would resolve some of the longstanding puzzles encountered in classical transfer theory, would generate a different set of predictions, and would therefore support very different experimental designs. The puzzle about why we see a lot of near transfer and not a lot of far transfer, for example, would be replaced with a set of specific research questions focusing on the ways the two CNDR transfer mechanisms interact to produce different amounts and qualities of transfer. One specific question would be how the effective "range" (or "receptive field" perhaps) of each mechanism relates to the amount of transfer it supports as the task contexts and contents become less similar.

A CNDR theory of transfer also suggests novel educational design principles. Assume, as I have suggested, that in a CNDR system context activates the relevant network of machines for a given task and task content provides the raw inputs to the network. There are many different ways a CNDR system could set up its network of machines (or neural circuits) during learning to solve any particular task. One goal of educational design is to shape educational outcomes across diverse learners more precisely (or at least more uniformly). One way to better control the learning outcome in

such a system would be to over-constrain the input—for instance, by providing additional inputs (visual, auditory, or tactile patterns unrelated to the actual content being learned) that would bias the system to configure its machine network in a particular way.

In a computerized chess training program, for example, to bias the student's learning process to self-organize around the principle of "material advantage," the computer could subtly (but perceptibly) manipulate the visual contrast or brightness of particular pieces or areas of the board to highlight configurations that systematically differentiate good and bad examples of material advantage. The student would presumably not have to be consciously aware of this feature for it to shape the internal representations that are being formed, as long as it is salient enough for the neural system to process it as an additional source of relevant information. One empirical question is whether the relevant knowledge would still be activated when the scaffolding is not present (for example, in a non-computerized game). If one wanted to bias the representations to organize around "control of the center" instead of "material advantage," then this same technique could be used to highlight board configurations that exemplify good and bad examples of that concept. This example is necessarily quite generic, as it is based on the most general features of the CNDR model. A more elaborated CNDR-based transfer theory could provide a more nuanced model of knowledge transfer that would both expand the range of educational design options and support more targeted interventions.

In the third step of the educational neuroscience research process, the prototype applications would be evaluated using standard methods—for example, controlled outcome studies. In an ideal world, the output from this step would be either a rigorously

evaluated and demonstrably effective educational application informed by neuroscience or some insight into why the design failed to produce the expected results that could be used to revise it.

The framework developed in this thesis is of course just one of many possibilities for applying neuroscience to education. Indeed, researchers in this domain draw on a wide variety of methods, including neuroimaging techniques like fMRI and PET (Goswami, 2004), methods like EEG and MEG that monitor the gross time course of neural processing (Dehaene, 1996), neural network models (Connell, 2002; Fischer & Connell, 2003), and dynamical systems models like mathematical growth models (Fischer & Bidell, 1998). One of the major obstacles researchers in this domain currently face, in fact, is the lack of a framework or even a common vocabulary for describing the standard methods in use and for mapping the relationships among these diverse approaches.

In an effort to contribute to current efforts to impose order on the emerging domain of educational neuroscience, in the remainder of this chapter I step back to discuss how the framework developed in this thesis fits into the larger universe of educational neuroscience methods. To make that possible, I first describe some features of the educational neuroscience domain that make it challenging to impose any kind of disciplinary order on it. I then suggest a framework for describing useful research methods based on the idea of a *pattern language* that has been successful in other domains with a similar structure, such as architecture and software design.

The Neuroscience-Education Domain

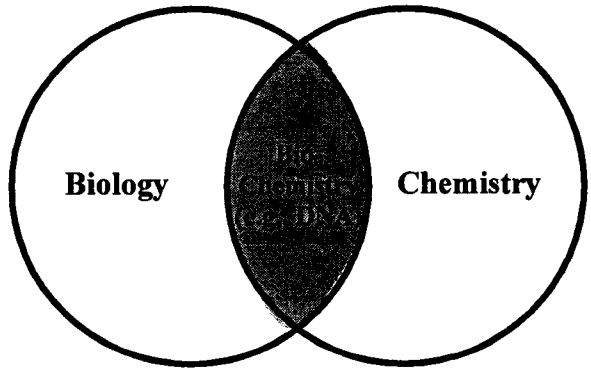
In the history of science, new "interdisciplinary" domains have sometimes emerged at the boundaries between established disciplines, where new and important problems arise that cannot be solved within the disciplines themselves but require new tools that combine and extend elements (theory, methods, tools) from the foundational disciplines. Biochemistry is one such example. The discovery of DNA—a set of chemical compounds that provide the blueprints for biological organisms—is the kind of event that requires the development of a new discipline (biochemistry) drawing on established disciplines (biology and chemistry) yet different from them (Figure 6.4a).

It is tempting to use an interdisciplinary domain like biochemistry as a reference model when thinking about the development of a new domain linking neuroscience to education (sometimes called "educational neuroscience" or "neuroeducation"). We might ask, for example, what problems lie at the intersection between neuroscience and education that could be fruitfully approached from a new hybrid discipline combining methods of both (Figure 6.4b).

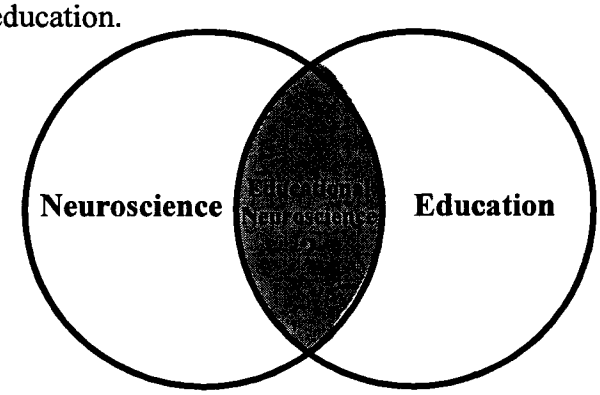
The problem with this analogy is that biology and chemistry are arguably true disciplines, each having its own paradigmatic theory and methods, but neither neuroscience nor education is a discipline in this same sense. On the one hand, "neuroscience" is a generic umbrella term covering many distinct disciplines, while on the other hand many educationists do not consider "education" to be a discipline at all in the technical sense, but instead a problem-based domain (Figure 6.4c) in which methods, data, and theory from a wide variety of disciplines are applied freely, often in *ad hoc* ways, to address specific problems arising in practice. Disciplines are defined by theory

Figure 6.4: Disciplines, interdisciplinary domains, and problem-based domains

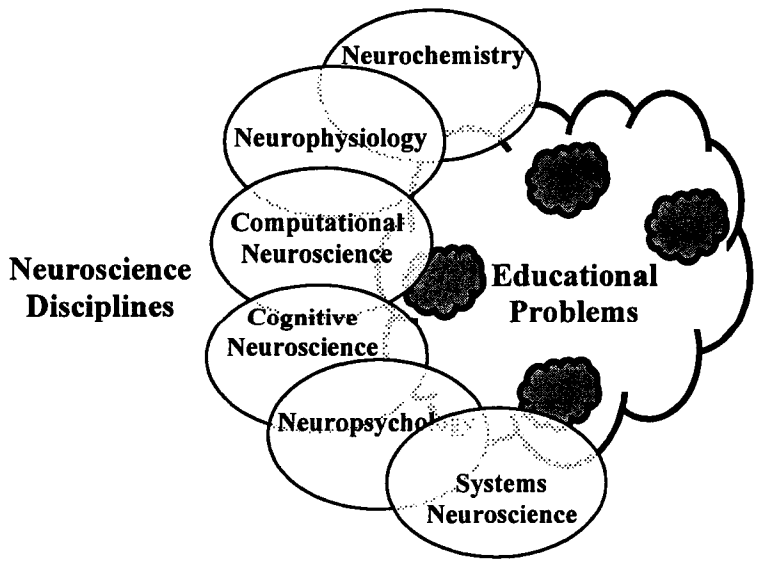
a) Biochemistry is an interdisciplinary domain bridging biology and chemistry.



b) Educational neuroscience seems like an interdisciplinary domain bridging neuroscience and education.



c) The difference is that neuroscience is not one but a set of disciplines, and education is not a discipline at all—it is a problem-based domain.



and methods, whereas education is defined by a context (or set of related contexts) and a set of problems.

In my view, we should not be thinking in terms of a "discipline" of educational neuroscience for the reasons just cited. Whatever it turns out to be, it does not look like an appropriate disciplinary candidate. In the absence of a disciplinary framework, the domain of educational neuroscience will obviously need some alternative structures to serve the same purposes. For example, disciplinary frameworks support theory development, guide the selection of research methods, support the evaluation and interpretation of results, and facilitate communication among members of the field. Organizational structures supporting goals like these would facilitate educational neuroscience research by providing standard templates for design, comparison, replication, and extension of research studies, and by helping researchers understand the diverse approaches being used, including their strengths, weaknesses, and inter-relationships.

The absence of a stable theoretical paradigm in educational neuroscience creates another set of challenges as well. That is, researchers within this emerging domain are responsible for conducting basic research (for example, conducing fMRI studies of dyslexia), facilitating the application of the basic research to education (for instance, determining how the fMRI results should inform diagnosis and remediation strategies), and evaluating the applications (e.g., assessing the accuracy of the diagnostic criteria and the associated interventions). It is highly unusual for all of these diverse requirements to be assigned to researchers within a single domain.

Consider the relationship between the theoretical domain of classical mechanics (within physics) and the applied domain of civil engineering, for example. Theoretical physicists do not typically design bridges—they develop theory (like Newton's laws of motion) that describes how some part of the world operates. In the case of Newton's laws, the theory has a very wide range of applicability. Similarly, civil engineers do not typically work out new physical laws—they mostly apply the relevant stable principles of physics (like Newton's laws) to solve specific problems (like designing cost-effective bridges that will not collapse). The relationship between classical mechanics and civil engineering is typical of the stable sciences and their sister applied disciplines (for example, electrical engineering is mainly an elaboration of Maxwell's equations, mechanical engineering is mainly grounded in Newton's laws and the fundamental laws of thermodynamics, chemical engineering is grounded in the theory of atomic interaction, etc.). In fact, it is reasonable to describe these applied domains (in their modern forms, anyway) as having grown out of the associated core theory or theories[1]. Electrical engineers do not solve arbitrary problems—they mostly limit their attention to problems that can be solved using applications of Maxwell's equations. Indeed, most of the literature in electrical engineering is a body of theory, research, and examples connecting Maxwell's abstract laws to specific contexts and problem types.

The relationship between neuroscience and education is very different from that between physics and civil engineering, mechanical engineering, or electrical engineering.

[1] This account is, of course, a simplification of the reciprocal process whereby disciplines are typically formed. One could argue, for example, that classical mechanics (Newton's laws in particular) grew out of efforts to solve applied problems in mechanical and/or civil engineering, which is also true enough (Gleick, 2004). I would suggest, however, that before Newton published his laws, mechanical engineering was not a true discipline in Kuhn's (1996) "paradigmatic" sense—it was a problem-based domain (as education is today, although probably much more coherent in the set of problems it sought to address). Newton's theoretical framework radically reorganized the domain from the foundation upward, providing the paradigmatic core that defines the discipline as we know it today.

On the one hand, neuroscientists have not identified general laws that are widely applicable to educational issues (Bruer, 2002). On the other hand, the diverse educational problems people seek to address through neuroscience have been around for a long time (much longer than modern neuroscience) and do not necessarily have much in common with each other. Instead of expanding a core theory outward to address a set of appropriate problems as in the more stable applied domains described above, the situation in educational neuroscience is reversed. That is, people are starting with an educational problem (like dyscalculia or dyslexia) and trying to "drill down" and solve it with "neuroscience" (writ large) without any clear indications at the outset of what specific tools, theory, or methods—if any—might be most appropriate (or even suitable).

Consider two areas that have been cited as models of promising work linking neuroscience and education: early mathematics (Bruer, 1997) and dyslexia (Pare-Blagoev, 2005). Based on cognitive developmental research, Case (1996) identified a "central conceptual structure" (CCS) for mathematics, which is a set of core cognitive structures (like the "mental number line" and its components) upon which a range of specific academic mathematical skills (such as basic arithmetic operations like addition and subtraction) depend. Based on this theory, Case and his colleagues developed a set of educational activities called *NumberWorlds* (Griffin, 2004; Griffin, Case, & Siegler, 1994) to facilitate the development of this mathematical CCS. Subsequent research has demonstrated the long-term efficacy of these interventions in supporting mathematical development for years following the intervention (Griffin, 2004; Griffin et al., 1994). In parallel with (and building on) the applied work of psychologists like Case, Griffin, and Siegler, cognitive neuroscientists (Dehaene, 1996, 1999; Dehaene, Spelke, Pinel,

Stanescu, & Tsivkin, 1999) have been mapping the neural correlates of the cognitive structures and processes supporting mathematical thinking (for example, the process of comparing two numbers to determine which is bigger). This kind of research helps to clarify the organization of cognitive processes like the mental number line that have been inferred from behavioral data—for example, by giving insight into whether the different processing steps occur in series or in parallel, and whether particular interventions target specific processing sites or whether their effects are more distributed and diffuse.

The study of dyslexia has followed a similar course. In one case, language researchers identified a specific deficit in the ability to process auditory inputs (for example, phonemes) of short duration that is implicated in a range of language-learning impairments such as dyslexia (Merzenich et al., 1996). Based on this and other findings, researchers created a computer-based intervention called *Fast ForWord* to remediate the condition (Morlet, Norman, Ray, & Berlin, 2003). In parallel with this applied work, cognitive neuroscientists have been conducting studies to map changes in patterns of neural activity that correlate with observed behavioral changes (Blake, Strata, Churchland, & Merzenich, 2002; Fiez, Raichle, Balota, Tallal, & Petersen, 1996; Fiez et al., 1995; Temple et al., 2003) to better understand the mechanisms involved (see Pare-Blagoev, 2005 for a comprehensive review).

In both cases of educational neuroscience research (mathematical development and dyslexia), the basic and applied research are being conducted simultaneously, and many researchers are participating directly in both aspects of the process. Moreover, the basic research findings tend to be applicable to a fairly narrow range of academic skills—not widely generalizable as in the case of Newton's laws—which means we are not yet

seeing the accumulation of stable generally applicable laws that would signal the emergence of a disciplinary core in educational neuroscience. For these reasons, it does not seem likely that the interdependence of basic and applied research in educational neuroscience studies is likely to change in the near future. These factors introduce special complexities into the domain and work against efforts to define it as a formal discipline.

Design Patterns

People working in other problem-based domains face some of the same challenges as educational neuroscience researchers—and for similar reasons. Architects and software design engineers, for example, must solve practical problems (e.g., design a skyscraper or automate a process, respectively) guided by general principles derived from best practices more than by any generally applicable core theory. In the absence of disciplinary frameworks, practitioners in both of these domains have benefited substantially by making use of a framework called *design patterns* (Alexander et al., 1977; Gamma, Helm, Johnson, & Vlissides, 1995) for identifying, abstracting, and formalizing elements of successful solutions to common problems for reuse. In the final pages of this dissertation, there is only space to sketch how design patterns would apply in educational neuroscience. In this section, therefore, I define what design patterns are and illustrate with a few concrete examples how they might be used to provide structure to this emerging hybrid domain.

Design Patterns Defined

Alexander and colleagues (Alexander et al., 1977), who are credited with originating the idea of design patterns in the domain of architecture, describe the basic

idea thus: "Each pattern describes a problem which occurs over and over again in our environment, and then describes the core of the solution to that problem, in such a way that you can use this solution a million times over, without ever doing it the same way twice" (p. *x*).

An architectural example of a recurring problem is the outdoor porch. Different porches serve different functions—for example, some porches are small and meant to shade entryways from the rain and snow, while others are large covered areas where people can sit outside shaded from the sun, and still others connect the interior of the building to a specific exterior space such as a courtyard. In addition, every porch design is unique. In terms of design patterns, what porches have in common (according to Alexander) is that they provide a transitional space that is neither inside nor outside, and these transitional spaces are important both practically (for example, to shelter people from the elements while waiting at the door) and psychologically (for instance, the transition from inside to outside or vice versa is less jarring if it is mediated by a space that has elements of interior spaces—like a roof—and exterior spaces—like open walls). Viewed in this way, the *Porch* design pattern provides much more useful information to support an architect than would a series of examples alone, because it specifies the criteria of a good porch design without constraining the specific details unnecessarily. The design pattern also formalizes and subjects to public scrutiny a set of well-defined criteria for distinguishing good designs from bad ones, which would otherwise only be implicitly defined in the heads of experts. Finally, by creating meaningful categories applicable to diverse exemplars to which simple names can be attached, design patterns

support the development of a common vocabulary to facilitate communication among members of the field.

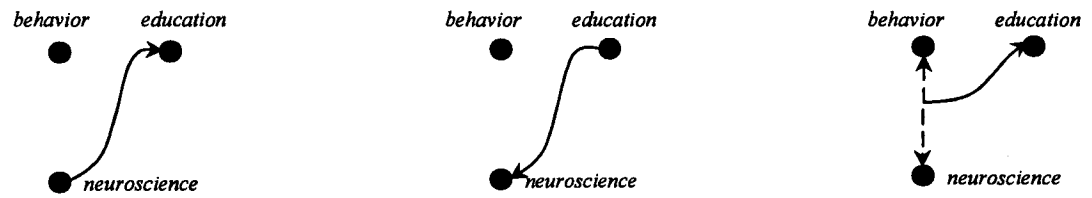
Examples of Design Patterns in Educational Neuroscience

I provide three sets of examples to demonstrate the potential utility of design patterns in educational neuroscience. The first set describes "false" design patterns. These patterns identify features that are common to many examples of bad educational neuroscience arguments and should generally be avoided. The second set illustrates how design patterns can facilitate the identification and naming of useful categories to highlight common patterns across diverse instances that appear very different on the surface. The third set of examples illustrates how the suppression of unnecessary detail through the use of design patterns emphasizes the meaningful differences between truly different patterns.

False Design Patterns

False design patterns identify recurring features of bad design (Figure 6.5), and should therefore be avoided. I discussed three such patterns in the introductory chapter of this thesis. The examples I use to illustrate these three patterns have all been discussed at length in the literature (Bruer, 1997, 1999a, 1999b, 2002); I use them here to exemplify common abstract patterns of reasoning. I have dubbed the first one the pattern of *Jumping to conclusions*. It is characterized by an unsupported leap from a particular observation about the brain to an educational recommendation without theoretical or empirical support. An example of this pattern is the jump from the observation that the two hemispheres of the brain appear to be functionally asymmetrical to the educational

Figure 6.5: Three common "false" design patterns in educational neuroscience



a) ***Jumping to conclusions*** b) ***Making data fit the theory*** c) ***Correlation, not causation***

conclusion that separate curricula should be tailored to each hemisphere ("right-brain/left-brain teaching"). There is no evidence to support this leap (Bruer, 1999a).

I call the second pattern *Making data fit the theory.* In instances of this pattern, people typically begin with a cherished educational conclusion and then trawl the neuroscience literature for findings that seem compatible with it. These isolated neuroscience facts are assembled into a narrative that seems to point toward the *a priori* conclusion. An example of this pattern is the notion that "education should be fun and stress-free." Starting from this "conclusion" (which is actually a premise), one might draw on research with rats suggesting that stressful conditions (via the amygdala) tend to narrow the animals' focus (i.e., toward the stressor and behaviors aimed at removing or avoiding it) compared to stress-free conditions in which they are more open, playful, and inquisitive. This finding might then be used as "evidence" in a chain of reasoning such as, "The brain's response to stress (via the amygdala) is to narrow focus, while the point of education is to broaden minds; therefore, educational environments should be stress-free, and even fun!" While many people would agree with the conclusion that education should involve low-threat environments and experiences, this is not a conclusion that follows from neuroscience—it is an *a priori* value arrived at through social consensus. In my experience, this pattern of *Making the data fit the theory* seems to be the one most favored by the brain-based education crowd, who take as their starting point a specific set of recommendations for educational reform and cite isolated findings from brain science to substantiate specific frameworks, programs, or interventions consistent with that agenda (see also Bruer, 1999a on this point). More generally, this pattern is the most useful to people who have something to sell, since they can start with the conclusion that

the product is good and work backwards to construct a narrative concerning how neuroscience indicates the need for the product.

The third false design pattern is called *Correlation, not causation.* This pattern is characterized by a pair of findings—one neurological and the other behavioral—that correlate in terms of time course, surface features, or some other dimension. This correlation is then used as the basis for an educational recommendation. For example, it has been observed that during the early years of life the brain sprouts many synapses (synaptogenesis), and that this period is followed by a period of massive pruning. It has also been observed that early in life children learn massive amounts of information rapidly (e.g., starting with no language they learn one in a couple of years), while many elderly people exhibit signs of mild to major cognitive decline (for example, including everything from minor general forgetfulness to major memory disorders like Alzheimer's disease). Based on the very loose correlation of the time dimensions of these two sets of events (one neurological and the other behavioral), people might conclude that the synaptogenesis and pruning processes are the cause of a slow cognitive decline starting in the early years and growing worse with each passing year until finally manifesting very noticeably in old age. The educational recommendation in this case might be to teach children as much as possible as early as possible ("use it or lose it!"). This example illustrates the false pattern of *Correlation, not causation* because the causal inference upon which the educational recommendation is made is not well supported by the observed correlation.

Together, these three abstract patterns account for many of the cases of fallacious reasoning in the popular press and in the "brain-based education" literature. My goal in

abstracting these false patterns is to help people (especially educational practitioners and non-educators in general) who wish to be critical consumers of the literature on neuroscience and education to distinguish valid arguments from spurious ones.

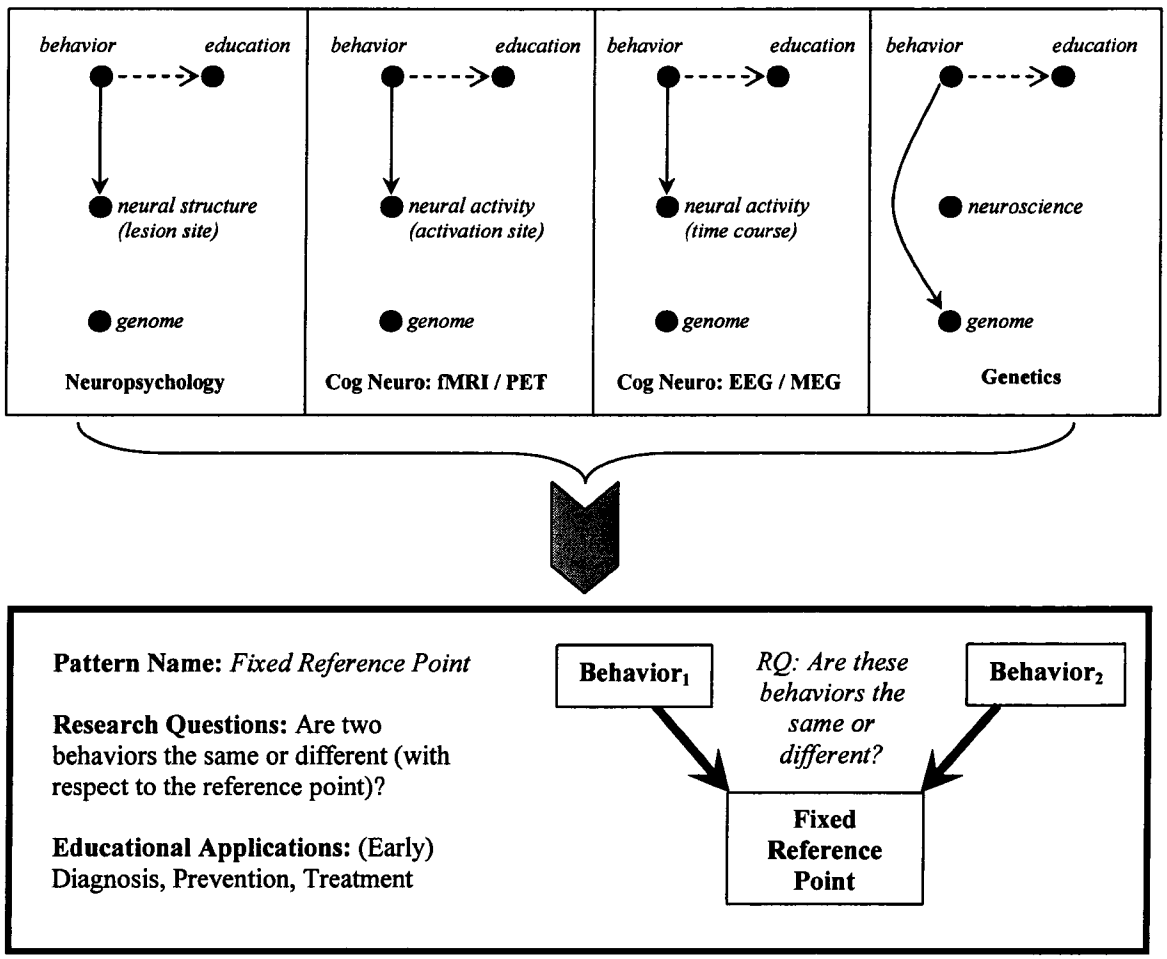
Design Patterns Suppress Details to Highlight Important Similarities

The second set of examples illustrates how a design pattern can unify a series of seemingly disparate instances (Figure 6.6). Consider four different sets of research methods (the lesion method, fMRI/PET, EEG/MEG, and genetic analysis) from three different disciplines (neuropsychology, cognitive neuroscience, and genetics) that provide information relevant to educational neuroscience:

1) Neuropsychologists correlate the location of structural brain damage (lesions) with patterns of atypical behavior as a way of inferring the functions performed in different areas of the brain (Banich, 1997). By providing information about how behavior changes when specific brain areas are selectively destroyed, this method provides information about where different functions are localized in the brain (for example, Broca's and Wernicke's language areas that are implicated in many language-related disorders) and how different areas depend on one another (for example, the role of the hippocampus in long term memory formation).

2) Cognitive neuroscientists use technologies like functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) that produce high spatial resolution pictures of brain activation to correlate specific behavioral patterns with the location of increased brain activity during a task. Researchers studying the effects of the *Fast ForWord* language training products have used

Figure 6.6: Four examples of the *Fixed Reference Point* design pattern in educational neuroscience



these kinds of technologies to study the neural correlates of behavioral changes induced by the intervention (for example, Temple et al., 2003).

3) Cognitive neuroscientists use technologies like electroencephalography (EEG) and magnetoencephalography (MEG) to correlate specific behavioral patterns with the time course of brain activity during a task. Researchers have used these techniques to identify the neural correlates of behaviors involved in mathematical reasoning (like number comparison), and to map the relationships among these processes (for example, Dehaene, 1996).

4) Geneticists use gene mapping techniques to correlate behavioral syndromes (e.g., autism) with specific genetic markers. In recent years, researchers have begun using these techniques to identify genetic markers associated with cognitive ability and learning disorders (Plomin & Walker, 2003).

In these four cases, the tools, techniques, and data are quite different from one another. The genetic example does not even involve the brain explicitly. What could they possibly have in common? Abstracting away from the details, the basic method in each case involves correlating a set of educationally relevant behaviors with some non-behavioral "fixed reference point," whether this is the location of a brain structure (e.g., a lesion), the location of brain activity (e.g., in fMRI), the time course of brain activity (e.g., in EEG), or the presence of specific genetic structures. For this reason, I have named the design pattern the *Fixed Reference Point* pattern.

Two key features of all of the methods exemplifying the *Fixed Reference Point* pattern are the following: 1) they are fundamentally comparative in nature, and 2) the data being compared is behavioral. In other words, instead of providing direct

information about what the brain or the genome is doing in any given case, these methods allow researchers to compare different behaviors to one another. Researchers often make inferences about the function of a specific brain or genome site based on these correlations, but that is simply a way of summarizing previous behavioral findings to facilitate future behavioral comparisons. For example, once Broca's area is identified from behavioral evidence as a "language area," then future experimentally elicited behaviors that are seen to depend on Broca's area are identified as "linguistic tasks" (or at least as tasks involving linguistic components). This shortcut allows researchers to implicitly compare one behavior to the whole history of behaviors studied previously (summarized via the label associated with the brain area), instead of having to run experiments involving multiple behaviors to make explicit comparisons every time.

Behavioral methods alone (such as those used in cognitive psychology) offer no comparable fixed reference point. Consequently, diagnostic categories based on behavioral data (such as "dyslexia") are often found upon closer inspection to aggregate many distinct sub-types exhibiting similar behavioral symptoms (e.g., "reading difficulties"). *Fixed Reference Point* methods can help differentiate such categories into meaningful sub-categories and even suggest potentially effective remediation strategies (e.g., central auditory problems are different from visual problems, the differences might not be easy to detect from behavior alone, and identifying which system is involved is useful in designing an intervention). In addition, these methods can identify potential linkages among behaviors that might seem dissimilar on the surface, if the different behaviors are found to involve overlapping brain centers. Therefore, one educational application for which these methods all seem very promising is early and precise

diagnosis, prevention, and treatment of learning disabilities (Plomin & Walker, 2003). The *Fixed Reference Point* design pattern seems to be the predominant one used in educational neuroscience at the present time.

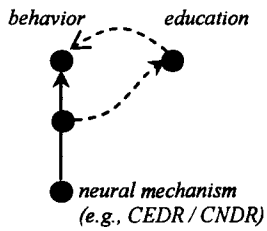
Design Patterns Suppress Details to Sharpen Important Distinctions

One of the primary benefits of design patterns is that they suppress unnecessary detail. In the previous example, I showed how suppressing detail can highlight meaningful similarities across different instances of the same pattern. In this final example, I discuss how suppressing superficial details can also help illuminate meaningful differences distinguishing fundamentally different patterns.

The *Fixed Reference Point* design pattern described in the previous section is grounded in behavior and projects "downward" onto the fixed point (a brain structure, brain activity, the genome, etc.). The first ("generation") step in the research framework summarized in Figure 6.1 projects in the opposite direction, from low-level neural mechanisms up to behavioral patterns that they cause. I have therefore dubbed this the *Bottom-up Mechanism* design pattern (Figure 6.7).

Abstracting away from the specific details of the framework itself (the use of computational models to generate predictions, the specific experimental methods used to validate the predictions, etc.), we are left with the generic research question motivating this step, which is about the causal relationship between a specific neural mechanism and one or more specific patterns of behavior. In my case study, I argued that this method (and therefore the design pattern it exemplifies) can suggest new neurally-grounded theoretical primitives that challenge existing behaviorally-grounded theories of educationally relevant phenomena like knowledge transfer. I expect that novel principles

Figure 6.7: The framework described in this thesis is one example of the *Bottom-up Mechanism* design pattern in educational neuroscience

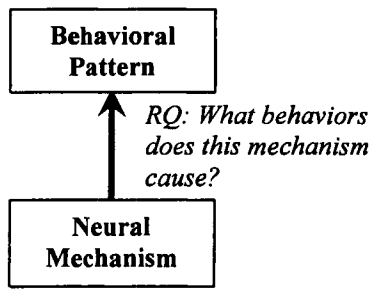


Connell framework

Pattern Name: *Bottom-up Mechanism*

Research Questions: What behavioral patterns are caused by (or associated with) the specified neural mechanism?

Educational Applications: Refined theories of educationally relevant behavior (e.g., knowledge transfer) grounded in neural mechanisms to inform research, curriculum design, and assessment.



will emerge from such research that can inform the design of pedagogical methods, curriculum materials, and assessment instruments.

A comparison of the *Fixed Reference Point* (Figure 6.6) and *Bottom-up Mechanism* (Figure 6.7) design patterns illustrates how meaningful differences are preserved across two distinct patterns even when the details are suppressed. In particular, the links between neural (or genomic) structures and behavior point in opposite directions in these two patterns. In addition, the kinds of research questions they can address and the kinds of educationally relevant insights they are likely to produce are quite different. The *Fixed Reference Point* pattern allows researchers to ask questions about similarities and differences between sets of behaviors as projected onto the reference point, and it provides information potentially useful for diagnosis, prevention, and remediation of atypical conditions. The *Bottom-up Mechanism* pattern allows researchers to ask questions about patterns of behavior that are caused by specific neural mechanisms, and it is more likely to provide information that would inform the design of more targeted interventions, assessments, and pedagogical methods for either the general population or special subgroups.

Note that these design patterns only pertain to the basic research step in each case. Separate design patterns should also be identified for the step of translating to education and evaluating the resulting intervention. Ideally, researchers should start building a library of patterns for each of the three steps that would facilitate each stage of research and design. New methods might be identified just by considering all combinatorial possibilities among the three sets of patterns. For example, using the *Bottom-up Mechanism* pattern for the basic research step, it might be possible to base the translation

(to educational designs) step on the neural mechanism itself (as long as some other validation method or at least a rigorous argument is provided so as to avoid *Jumping to Conclusions*), on the behavioral pattern(s) caused by the neural mechanism, or on the relationship between the two. Each of these possibilities might be described by a distinct translation design pattern that could also be used in conjunction with other basic research design patterns and a variety of evaluation design patterns.

(Neuro)Scientific Research in Education

In recent years, researchers, policy makers, funding agencies, and the federal government (among others) have been calling for more "scientifically based research in education" (Eisenhart & DeHaan, 2005; National Research Council, 2002). It is difficult to define "scientific research" even in the natural sciences. It is much more difficult to define what it means to do scientific research in the social sciences and education (Eisenhart & DeHaan, 2005). In an effort to provide some standards in this area that are reasonably uniform yet flexible enough to apply to the diverse range of quantitative and qualitative methods being used in education, the National Research Council has identified six characteristics associated with scientific research (see Table 6.1; I added the short labels for easy reference).

These guiding principles seem to be offered more in the spirit of a set of "symptoms" of scientific research in education rather than as a necessary and sufficient checklist. In other words, a research study does not necessarily have to exhibit all of these properties to the maximal degree all the time to be considered scientific. Conversely, a study exhibiting all of these symptoms can utterly fail to be scientific (for example, if the "theory" it draws on is not scientific). In general, however, studies

exhibiting more of these features (and/or treating each feature more thoroughly) will tend to be more scientific than studies exhibiting fewer features (and/or treating individual features less thoroughly). Note, however, that these principles will apply differently and their relative importance will be weighted differently depending on the context(s), discipline(s), and method(s) involved in any particular case.

Table 6.1: Six "guiding principles" for scientific research in education

Scientific research in education...	
...poses significant questions that can be investigated empirically	(*Relevance, Tractability*)
...links research to relevant theory	(*Cumulativity*)
...uses methods that permit direct investigation of the question	(*Validity*)
...provides an explicit and coherent chain of reasoning	(*Soundness*)
...replicates and generalizes across studies	(*Robustness, Generality*)
...makes research public to encourage professional scrutiny and critique	(*Transparency*)

(Eisenhart & DeHaan, 2005, p. 3; National Research Council, 2002, pp. 3-5, 54-72)

The guiding principles for scientific research do not themselves constitute a design pattern, but they can be used productively in conjunction with the kinds of design patterns described in previous sections. In particular, the symptoms of scientific research in Table 6.1 represent one set of normative standards that can be used to evaluate the quality of design patterns in educational neuroscience, to identify their specific weaknesses, and to compare and contrast different design patterns to one another along an evaluative dimension pertaining to scientific rigor. The benefit of applying this set of standards to abstract design patterns instead of (or in addition to) specific studies is that any conclusions drawn about a design pattern in this regard can be applied to all specific instances derived from the abstract pattern. In other words, the design patterns provide a

mechanism whereby lessons learned and conclusions drawn about the scientific merits and weaknesses in one instance (a specific research study) can be highly leveraged through propagation to many—if not all—other instances of the same pattern.

A research design or argument can fall short of scientific standards by failing on one or more of the dimensions described in Table 6.1. The three "false" design patterns described previously, for example, fail primarily and most clearly on the soundness dimension. These instances of these patterns completely fail to provide an explicit and coherent chain of reasoning linking neuroscience findings to educational implications and instead jump directly from neuroscience "premises" to educational "conclusions" (or vice versa). In addition, without an explicit and coherent argument linking premises to conclusions, there is no way to establish validity, transparency, cumulativity, robustness, or generality. In fact, the only dimension on which instances of the false design patterns might succeed is relevance (although it is not clear the questions being addressed are tractable).

In terms of logical structure, the three false design patterns in educational neuroscience (*Jumping to conclusions, Making the data fit the theory,* and *Correlation, not causation*) are distinct from each other. When viewed in the light of criteria for scientific research, however, their common flaw becomes evident: they lack a coherent argument linking neuroscience findings to educational applications. In many specific instances of these patterns there is a bona fide scientific finding from neuroscience at the core of the argument. The problem arises when people try to claim that the entire argument (from neuroscience to education) is scientific because one or more of its premises are. Just because a finding has been generated using scientific methods,

however, does not mean that any arguments based on it or that any conclusions drawn from it are scientific. The output (conclusion) of one phase of research is the input (hypothesis) to the next, and the six guiding principles need to be applied to each phase of a study and/or argument. I have argued that there are at least three phases that need to be handled in this way in applying basic neuroscience research to educational practice. Arguments based on the false design patterns generally involve at most one scientifically rigorous link, and they often erroneously attempt to stretch the scientific "halo" of that link to cover the larger argument in its entirety.

The examples of the *Fixed Reference Point* pattern I described previously (i.e., studies based on the lesion method, fMRI, EEG, and genetic studies) meet all six criteria of scientific research, but only with respect to the first step (basic research) in the three-step research framework depicted in Figure 6.1. *Fast ForWord* and *NumberWorlds* are two examples that build on the basic research step by adding application and evaluation links. The evaluation links in these cases (which take the form of outcome studies evaluating the effectiveness of the interventions) support all six criteria as well. The application link seems to be of a different sort, however—it requires a chain of reasoning linking neuroscience findings to educational designs, but it depends on the basic research and evaluation steps for some of the other features (like validity). In other words, the application step seems to be of a qualitatively different nature (with respect to the criteria of science) than the basic research and evaluation steps, which—although very different in the details—have much in common with each other at an abstract level in terms of the guiding principles in Table 6.1.

These examples suggest that scientific research in education entails some special sources of complexity, especially as compared to scientific research in the natural sciences. In particular, "scientific research" has more than one meaning in education, whereas in any given scientific or engineering discipline its meaning is much more narrowly and uniformly defined. Furthermore, the guiding principles in Table 6.1 might have to be applied multiple times in qualitatively different ways in a single educational study. For example, the principles need to be applied appropriately to each of the three steps in the educational neuroscience research framework (Figure 6.1), and specific requirements and procedures will differ by discipline, level of analysis, basic research vs. application vs. evaluation step, etc. Applying the principles once (for example in the basic research step within neuroscience) and then drawing *ad hoc* conclusions about educational practice (as instances of the false design patterns do) does not constitute a scientific argument.

As I mentioned previously, there are few—if any—examples of scientific research in educational neuroscience that are fully documented in the literature. This observation does not imply that researchers are not privately drawing on neuroscience to inform their educational designs in authentic and meaningful ways. It does mean, however, that anyone proceeding in that manner fails on the transparency dimension, and therefore such work does not meet one of the key criteria of scientific research. Science is an inherently public process. Even if such work were scientific in every other regard, if the reasoning process is not made public then the work's validity, soundness, robustness, generality, etc. cannot be evaluated by qualified members of the field, and the

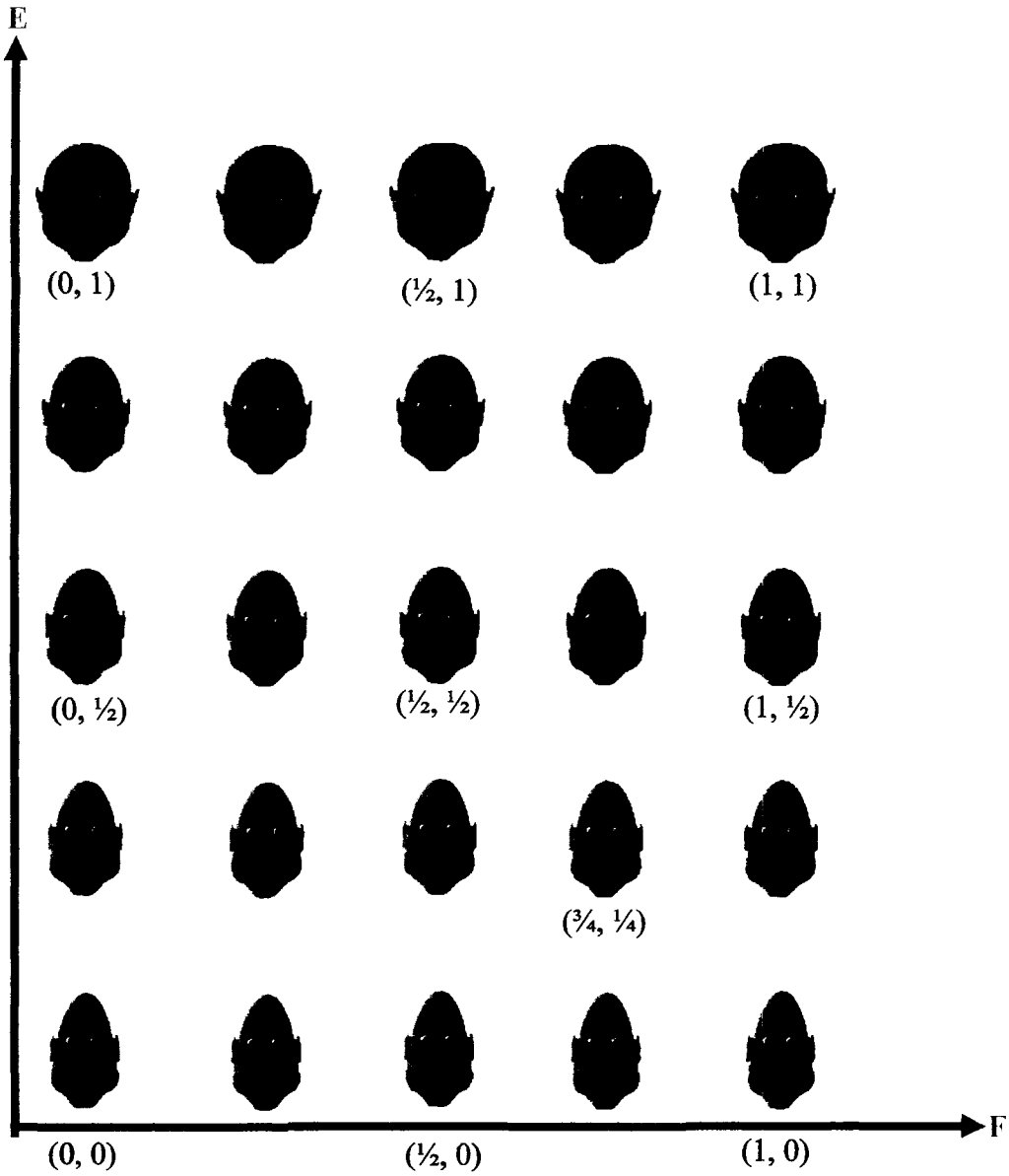
work does not support the accumulation of robust, publicly accessible knowledge that is one of the hallmarks of scientific research.

In this dissertation, I have sought to provide a framework, a set of tools, and a case study of educational neuroscience research. I believe the research meets (or at least supports) all six criteria described in Table 6.1. Regardless of whether my arguments and analyses turn out to be substantively right or wrong, I hope this study will support discussion and debate by serving as a concrete example of scientific research in education conducted specifically within the context of a school of education. Indeed, given the interdisciplinary and application-oriented nature of the work, I do not know where else I could have successfully carried it out.

Appendix A

Experimental Stimuli

The "face-space" used in the experiment, with representative points labeled



I generated the face stimuli using the software package *Poser 5* by Curious Labs, Inc. (http://www.curiouslabs.com). I made the prototype face (in the center) by modifying a stock face to make it look more alien. I then systematically varied the amount of head tapering to change the distance between the eyes (E), and I varied the distance from mouth to eyes to change the face height (F) in equal increments to generate all of the other faces from the prototype.

Appendix B

Recruiting Flyer

Research Subjects Needed for a Study of Cognition

I am currently recruiting subjects over the age of 21 to participate in a cognitive study as part of my dissertation research. The purpose of the study is to test some predictions about knowledge construction derived from a computer simulation of human cognition. Participants will be asked to complete a task involving visual discrimination and categorization of faces.

Duration: Approximately 60 minutes
Compensation: $10

Contact: Michael Connell
Michael_Connell@gse.harvard.edu

Appendix C

Informed Consent Form

Informed Consent Form

Please read this consent agreement carefully before agreeing to participate in this experiment.

Purpose of the experiment:
To examine the process of knowledge construction.

What you will do in this experiment:
You will view a series of faces on a computer display and categorize them into two groups using two buttons on the computer keyboard or other input device (such as a game controller). Periodically you will make judgments about which of two faces is most similar to a third face, using two other buttons to indicate your response. After you complete the computerized tasks, you will answer some questions about your experience in the experiment. If you give your permission, I will audiotape your responses to these questions.

Time required:
The experiment will take approximately 60 minutes to complete.

Risks:
There are no anticipated risks associated with participating in this study. The effects of participating should be comparable to those you would experience from viewing a computer monitor for 60 minutes and using a keyboard or game controller.

Benefits:
You will receive $10 for participating in this experiment. At the end of the experiment, I will provide a thorough explanation of the experiment and of my hypotheses. I will describe the potential implications of the results of the study both if my hypotheses are supported and if they are disconfirmed. I will also be happy to answer any questions you might have about this research.

Confidentiality:
Your participation in this experiment will remain confidential, and your identity will not be stored with your data. Your responses will be assigned a code number, and the list connecting your name with this number will be kept in a secure location and will be destroyed once all the data have been collected and analyzed. If you consent to having your post-test responses audiotaped, the tapes will be transcribed, identifying information will be removed from the transcript, and the transcript will be labeled with the same identifying number as the other data. The audiotapes will be stored in a secure location and erased once all the data have been collected and analyzed.

Future Uses of Experimental Data:
Your response data and any transcripts, stripped of personal identifiers, may be included anonymously in future analyses, demonstrations, or presentations.

Participation and Withdrawal:
Your participation in this experiment is completely voluntary, and you may withdraw from the experiment at any time without penalty. You will receive payment based on the proportion of the experiment that you completed. You may withdraw by informing the experimenter that you no longer wish to participate (no questions will be asked).

Contact:
If you have questions about this experiment, please contact Michael W. Connell at the Graduate School of Education, Harvard University via email: Michael_Connell@gse.harvard.edu.

Who to contact about your rights in this experiment:
Jane Calhoun, Harvard University Committee on the Use of Human Subjects in Research, Science Center 128, Cambridge, MA 02138. Phone: 617-495-5459.
E-mail: jcalhoun@fas.harvard.edu.

Agreement:
The purpose and nature of this research have been sufficiently explained and I agree to participate in this study. I understand that I am free to withdraw at any time without incurring any penalty.

Signature: _____ Date: _____

Name (print): _____

Appendix D

Background Questionnaire

Background Information

Age (years and months):
Sex (circle one): Male Female
Computer Experience: a) Number of years you have owned a personal computer: _____ b) Number of hours per week you have spent using a computer during the past twelve months (on average): _____

Appendix E

Taxonomy of Nested Multi-level Regression Models for Categorization Task

Table E.1: Taxonomy of fitted linear multilevel models describing the relationship between ln(reaction time) in a dichotomous categorization task and the distance of the stimulus from the category boundary, controlling for subject's age, sex, computer experience hours and years), as well as the interactions between stimulus distance and these control variables (subjects=48, observations=959)

	Model								
Predictor	**M0t (uncond)**	**M1t**	**M2t**	**M3t**	**M4t**	**M5t**	**M6t**		**M7t_a**
Intercept	6.6438****	6.7285****	6.7277****	6.6881****	6.7559****	6.719****	6.6995****		6.6896****
DIST		-0.0845****	0.08451****	0.06542****	0.09808****	0.08032****	0.07093****		-0.04957**
FEMALE			-0.06347	0.009776	-0.06349	-0.06338	-0.06349		0.0244
AGE			0.000186	0.000186	0.00002	0.000186	0.000186		
COMP_YRS			-0.00617	-0.00617	-0.00617	-0.00175	-0.00617		
COMP_HRS			0.001257	0.001256	0.001257	0.001259	0.003526		0.002066
DIST*FEMALE				-0.03527~					-0.03795~
DIST*AGE					0.00008				
DIST*COMP_YRS						-0.00213			
DIST*COMP_HRS							-0.00109*		-0.00116*
σ_{u0}^2	0.03769****	0.04624****	0.0487****	0.04868****	0.04869****	0.04865****	0.04866****		0.04825****
σ_{u0u1}		-0.00369	-0.00491~	-0.00489~	-0.0049~	-0.00489~	-0.00489~		-0.00434~
σ_{u1}^2		0	0	0	0	0	0		0
σ_{ε}^2	0.09863****	0.09112****	0.09112****	0.09079****	0.09097****	0.09094****	0.0907****		0.09032****
-2LL	603.6	528.9	526.1	522.8	524.6	524.3	521.9		520.3

Key: ~ p<.10; * p<.05; ** p<.01; *** p<.001; **** p<.0001

Table E.1 (cont'd): Taxonomy of fitted linear multilevel models describing the relationship between ln(reaction time) in a dichotomous categorization task and the distance of the stimulus from the category boundary, controlling for subject's age, sex, computer experience hours and years), as well as the interactions between stimulus distance and these control variables (subjects=48, observations=959)

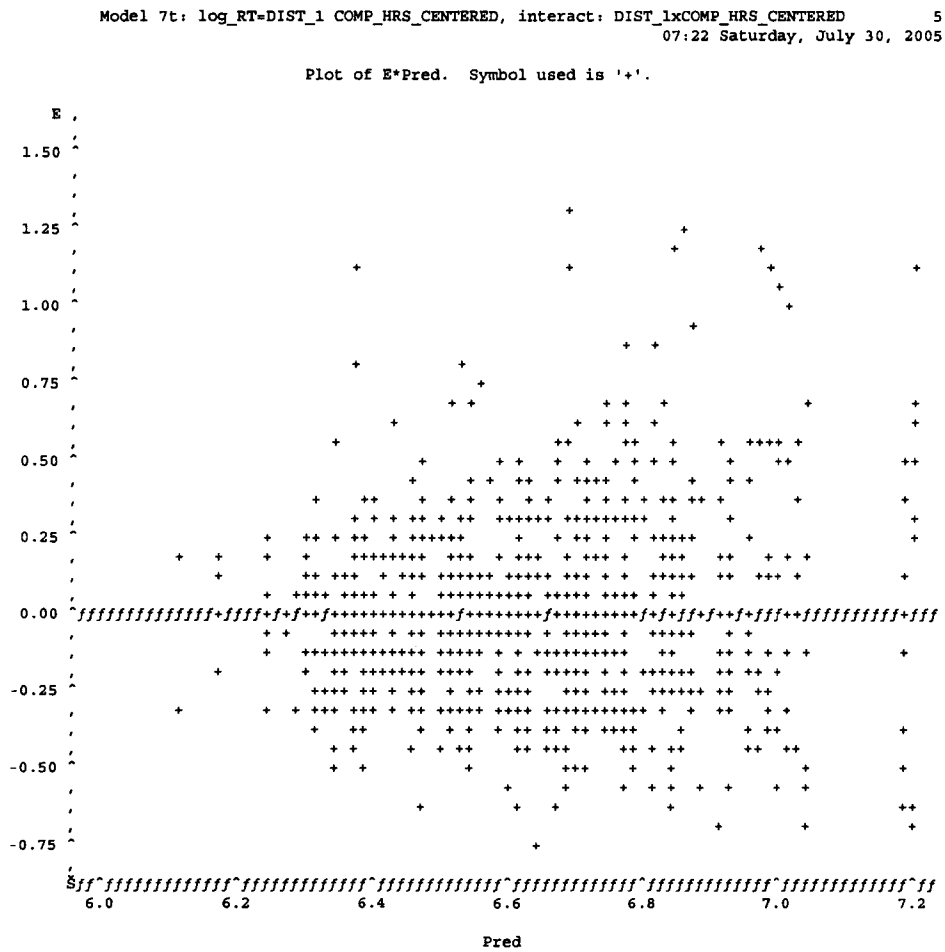
	Model							
Predictor	**M7t_b**	**M7t_c**	**M7t_d**	**M7t_e**	**M7t_e2**	**M7t_f**	**M8t**	**M9t**
Intercept	6.6511****	6.6431****	6.7315****	6.7037****	6.7039****	6.7315****	6.7179****	6.6896****
DIST	-0.07531***	-0.07093****	-0.07093****	-0.07079****	-0.07093****	-0.07093****	-0.06542****	-0.04957**
FEMALE			-0.04999			-0.04999	0.01961	0.002066
AGE	0.00024	0.000277						
COMP_YRS				-0.00233	-0.00395			
COMP_HRS	0.002946	0.003089	0.001937	0.002361	0.002596	0.001937		0.0244
DIST*FEMALE							-0.03527~	-0.00116~
DIST*AGE	0.00002							
DIST*COMP_YRS				-0.00088				
DIST*COMP_HRS	-0.00101~	-0.00109*	-0.00109*	-0.00096	-0.00109*	-0.00109*		-0.03795*
σ_{u0}^2	0.04537****	0.04537****	0.04825****	0.04624****	0.04625****	0.04825****	0.04784****	0.04825****
σ_{u0u1}	-0.00379	-0.00379	-0.00435	-0.0038	-0.00381	-0.00435~	-0.00425~	-0.00434~
σ_{u1}^2	0	0	0	0	0	0	0	0
σ_{ε}^2	0.09069****	0.0907****	0.0907****	0.09067****	0.0907****	0.0907****	0.09079****	0.09032****
-2LL	523.1	523.2	524	523.9	524.1	524	525	520.3

Key: ~ p<.10; * p<.05; ** p<.01; *** p<.001; **** p<.0001

Appendix F

Residuals for Final Model (Categorization Task)

Level-1 residuals: Categorization Task



NOTE: 1 obs had missing values. 420 obs hidden.

Level-1 Residuals: Categorization Task (cont'd)

Model 7t: log_RT=DIST_1 COMP_HRS_CENTERED, interact: DIST_1xCOMP_HRS_CENTERED 8
07:22 Saturday, July 30, 2005

The UNIVARIATE Procedure
Variable: E

Histogram	#	Boxplot
1.25+*	2	*
.**	5	0
1.05+*	2	0
.*	2	0
0.85+*	3	0
.**	5	0
0.65+***	11	0
.*******	26	|
0.45+*******	25	|
.***************	56	|
0.25+******************	65	|
.************************	100	+-----+
0.05+************************************	144	| + |
.**	153	*-----*
-0.15+*********************************	127	+-----+
.*****************************	102	|
-0.35+*****************	64	|
.***********	43	|
-0.55+****	15	|
.**	7	|
-0.75+*	2	0

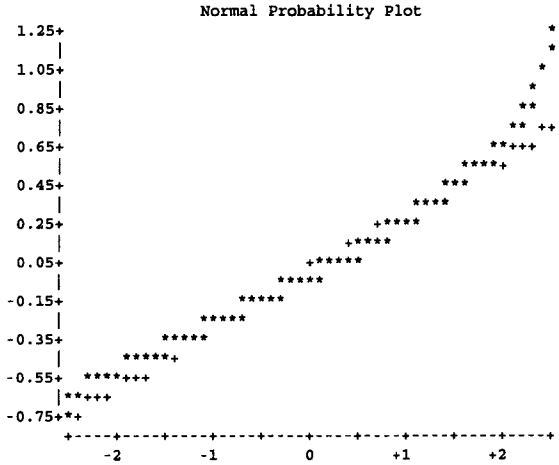
* may represent up to 4 counts

Level-1 Residuals: Categorization Task

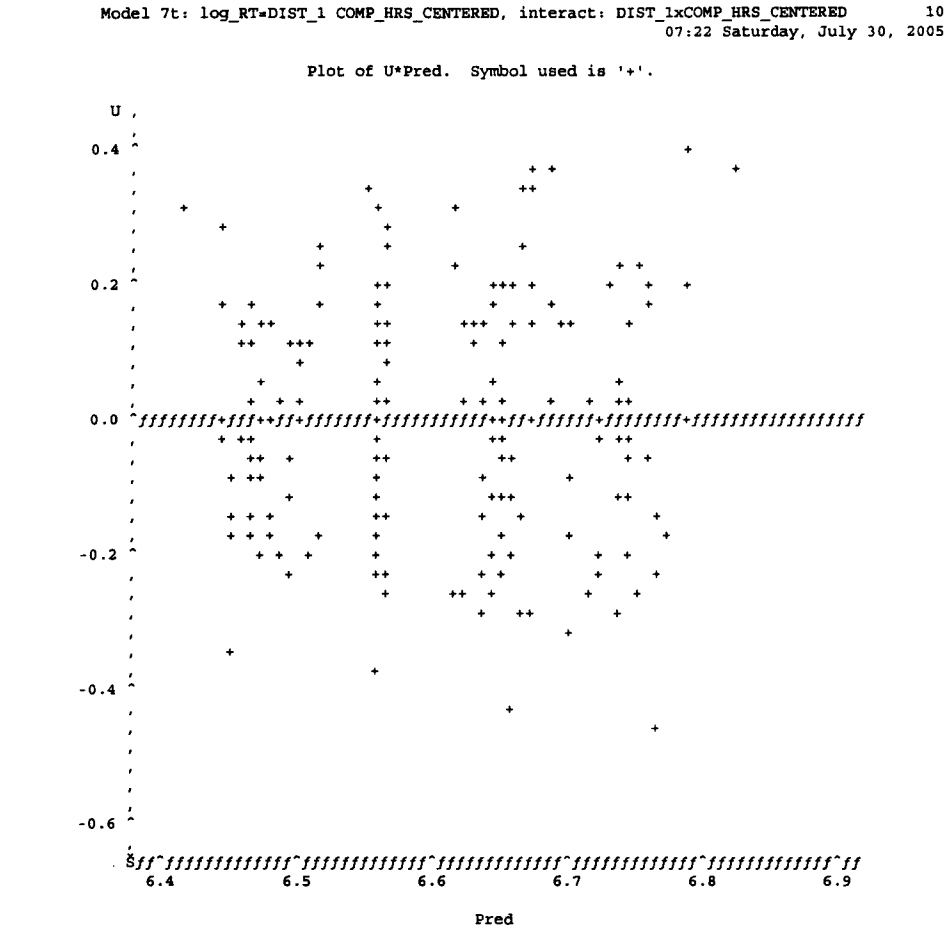
Model 7t: log_RT=DIST_1 COMP_HRS_CENTERED, interact: DIST_1xCOMP_HRS_CENTERED 9
07:22 Saturday, July 30, 2005

The UNIVARIATE Procedure
Variable: E

Normal Probability Plot



Level-2 Residuals: Categorization Task



NOTE: 1 obs had missing values. 797 obs hidden.

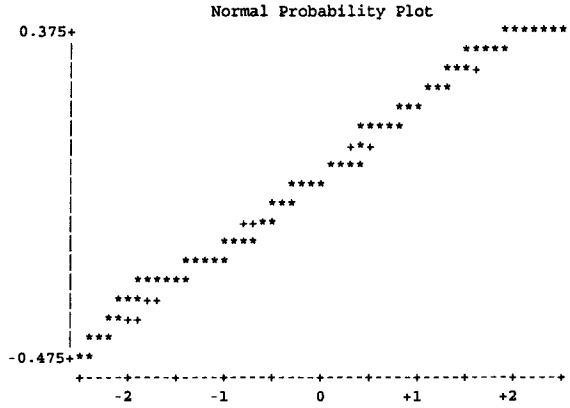
Level-2 Residuals: Categorization Task (cont'd)

Model 7t: log_RT=DIST_1 COMP_HRS_CENTERED, interact: DIST_1xCOMP_HRS_CENTERED 13
07:22 Saturday, July 30, 2005

The UNIVARIATE Procedure
Variable: U

Histogram	#	Boxplot
0.375+**********	28	
.**************	38	
.*******	20	
.*******************	55	
.************************	72	
.**	120	+-----+
.**	6	| |
.**	120	| |
.**	134	*--+--*
.*********************	60	| |
.********************	58	| |
.********************************	88	+-----+
.***************************	78	|
.******************	54	|
.****	10	|
.**	4	|
.**	6	|
-0.475+***	8	|

* may represent up to 3 counts



Appendix G

Taxonomy of Nested Multi-level Regression Models for Similarity Task

Table G.1: Similarity Analysis. Taxonomy of fitted mixed logistic models describing the relationship between fraction of within-category pairs selected in a visual similarity judgment task and time (while learning was taking place) controlling for subject's age, sex, computer experience (hours and years), as well as the interactions between time and these control variables (subjects=48, observations=240).

	Model						
Predictor	**MUM (uncond means)**	**M0 (uncond growth)**	**M1**	**M2**	**M3**	**M4**	**M4a**
Intercept	0.1746**	-0.04474	-0.04177	[illegible]	-0.03141	0.1196	-0.1219
Time		0.1192**	0.1316*	[illegible]	0.06881	0.04789	0.3535**
Female			-0.00532				
Age				[illegible]			0.000399
Comp_Yrs					-0.00111		
Comp_Hrs						-0.00507~	-0.0034
Time*Female			-0.02327				
Time*Age				[illegible]			-0.0005*
Time*Comp_Yrs					0.004201		
Time*Comp_Hrs						0.0022	
σ_{u0}^2	0.1514***	0.04355	0.04349	[illegible]	0.04317	0.03446	0.03218
σ_{u0u1}		-0.00099	-0.00097	[illegible]	-0.00055	0.003002	0.005849
σ_{u1}^2		0.04049**	0.04018**	[illegible]	0.03958**	0.03886**	0.03546**
σ_{ε}^2	0.01446****	0.008634****	0.008641****	[illegible]	0.008645****	0.008639****	0.008617****
-2LL	-268.9	-338.3	-338.4	[illegible]	-338.9	-341.9	-345.6

Key: ~ p<.1; * p<.05; ** p<.01; *** p<.001; **** p<.0001

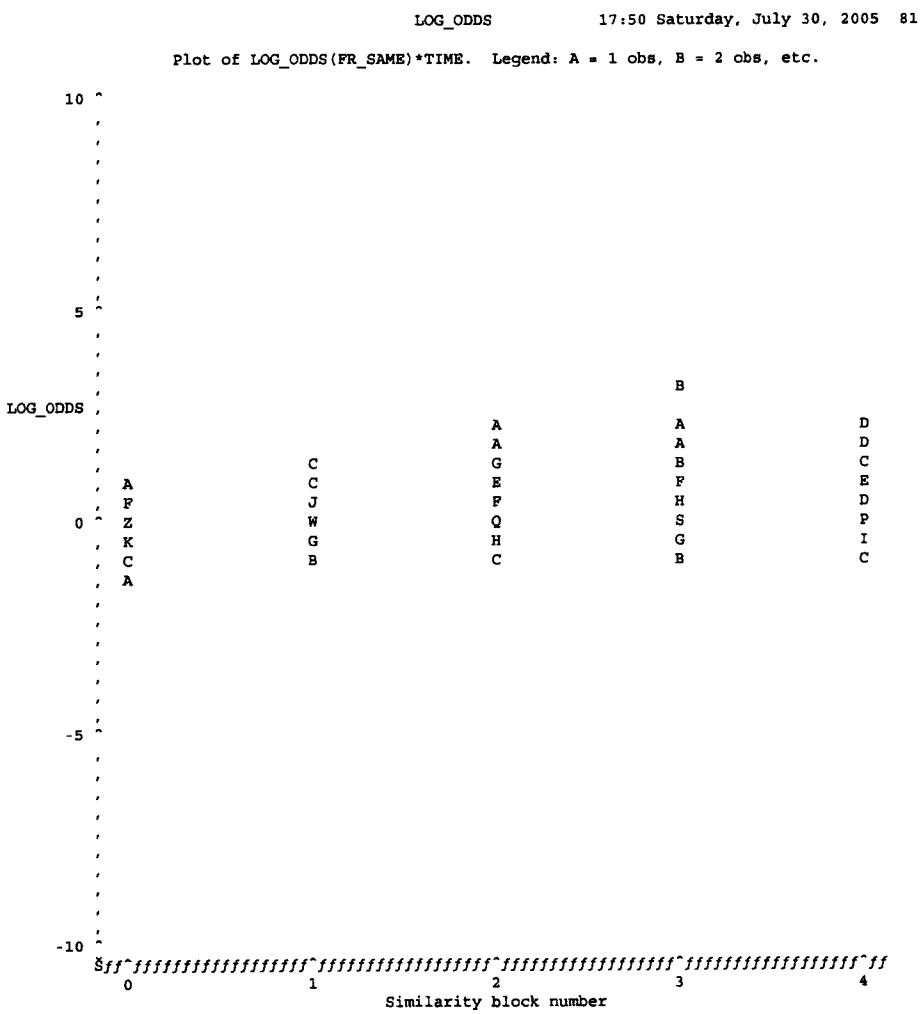
Table G.1: (cont'd) Similarity Analysis. Taxonomy of fitted mixed logistic models describing the relationship between fraction of within-category pairs selected in a visual similarity judgment task and time (while learning was taking place) controlling for subject's age, sex, computer experience (hours and years), as well as the interactions between time and these control variables (subjects=48, observations=240).

	Model	
Predictor	**M4b**	**M4c**
Intercept	0.08462	-0.1118
Time	0.1195*	0.3316~
Female		
Age		0.000389
Comp_Yrs		
Comp_Hrs	-0.00399	-0.00357
Time*Female		
Time*Age		-0.00048~
Time*Comp_Yrs		
Time*Comp_Hrs		3.63E-04
σ_{u0}^2	0.03498	0.03215
σ_{u0u1}	0.001991	0.005901
σ_{u1}^2	0.0407*	0.03539*
σ_{ε}^2	0.008634****	0.008618****
-2LL	-340.5	-345.6

Key: ~ p<.1; * p<.05; ** p<.01; *** p<.001; **** p<.0001

Appendix H

Logistic Assumption Verification (Similarity Task)



Appendix I

Post-Experimental Questionnaire

Please answer the following questions about the faces you saw in the experiment.
May I audiotape your responses to these questions? YES NO Initials _____

1) **Written Verbal Description**

Briefly describe (in words) what a prototypical **GORF** looks like:

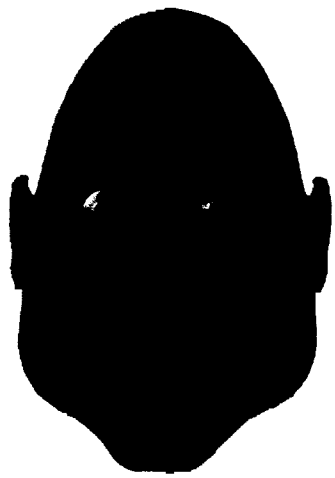
Briefly describe (in words) what a prototypical **DIMP** looks like:

2) **Talk Aloud**

During the experiment, when you were categorizing a face like the one shown below, what was your strategy for deciding whether it is a Dimp or a Gorf? (Describe your reasoning process out loud). In particular:

- What features helped you decide what kind of creature it is?

- Did you rely on any explicit rules to make the decision? If so, what rules?



What kind of creature is this? (Circle one) GORF DIMP

3) **Scaffolded Talk-Aloud**

Looking at the sets of GORFs and DIMPs on the next page, please answer the following questions, verbalizing your thinking aloud.

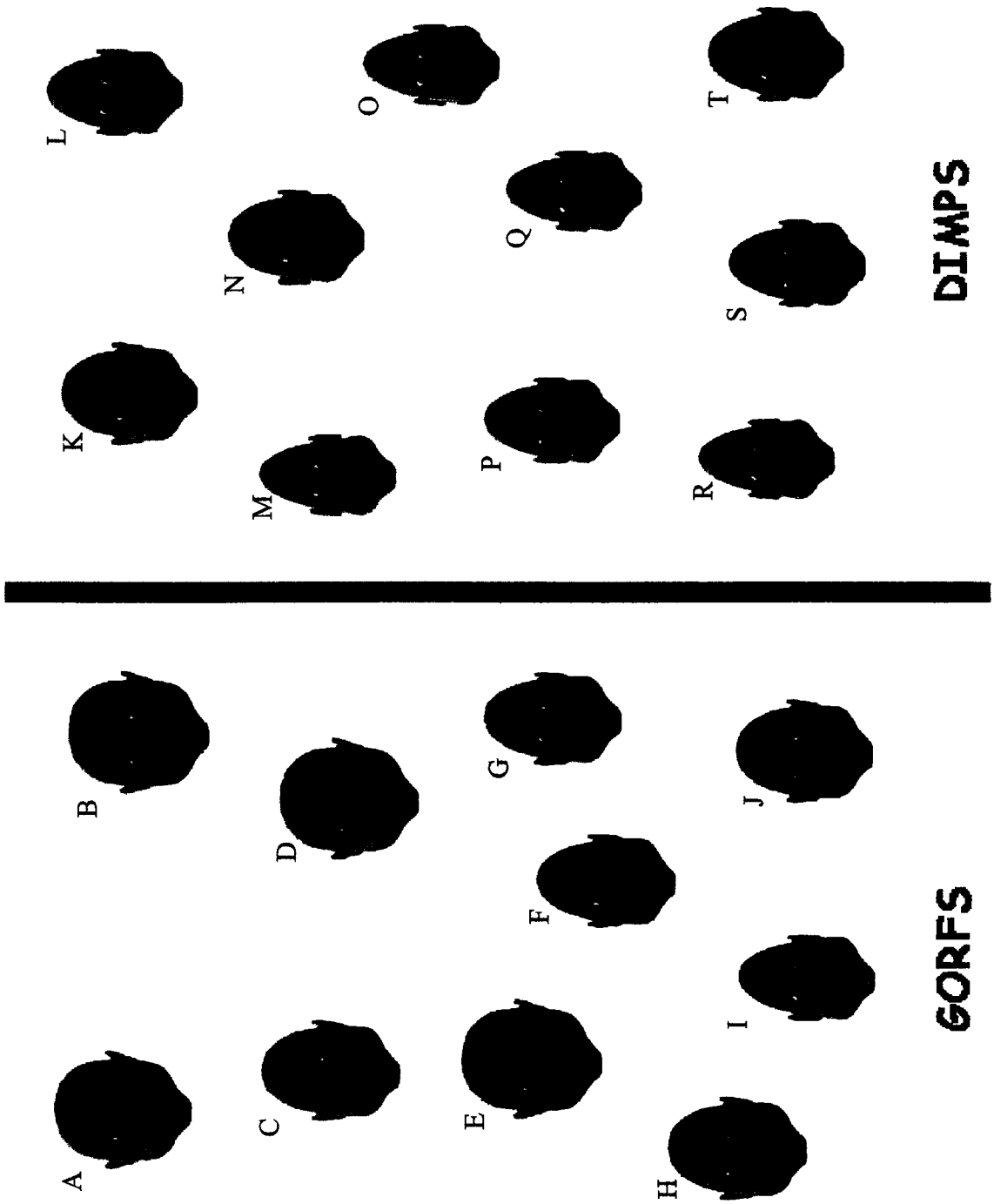
a) Circle two GORFs and two DIMPs that were relatively easy for you to identify (if there were any). What strategy did you use to identify them during the experiment, and what made them easy?

b) Put an "X" through two GORFs and two DIMPs that you found relatively difficult to identify (if there were any). What strategy did you use to identify them during the experiment, and what made them difficult?

c) Looking at the two sets as a whole, describe the strategy/strategies you used to categorize them during the experiment by first describing a feature or rule that you used (below) and then identifying specific faces to which you applied it by labeling them with the rule number (R1, R2, etc.).

R1:

(Add as many rules as necessary below)



References

Abbott, L., & Sejnowski, T. J. (Eds.). (1999). *Neural codes and distributed representations: Foundations of neural computation.* Cambridge, MA: The MIT Press.

Abdi, H., & Valentin, D. (1994). Neural, connectionist, and numerical models of face recognition. *Psychologie Francaise, 39*(4), 375-391.

Abu-Rabia, S., & Kehat, S. (2004). The critical period for second language pronunciation: Is there such a thing? Ten case studies of late starters who attained a native-like hebrew accent. *Educational Psychology, 24*(1), 77-98.

Addanki, S. (1984). *Applications of connectionist modeling techniques to simulations of motor control systems.* University of Rochester (Unpublished Thesis), New York.

Alberts, B., Bray, D., Lewis, J., Raff, R., Roberts, K., & Watson, J. D. (1994). *Molecular biology of the cell* (Third ed.). New York: Garland.

Alden, B. E., & Bramer, M. A. (1988). An expert system for solving retrograde-analysis chess problems. *International Journal of Man-Machine Studies, 29*(2), 97-112.

Alexander, C., Ishikawa, S., Silverstein, M., Jacobson, M., Fiksdahl-King, I., & Angel, S. (1977). *A pattern language.* New York: Oxford University Press.

Altmann, E. M., & Burns, B. D. (2005). Streak biases in decision making: Data and a memory model. *Cognitive Systems Research, 6*(1), 5-16.

Andersen, F. O. (1999). Neural networks: Reading, spelling and dyslexia. *Psykologisk Paedagogisk Radgivning, 36*(1), 38-51.

Andersen, O. S., & Koeppe, R. E. (1992). Molecular determinants of channel function. *Physiological Review, 72*, S89-S158.

Anderson, B., & Donaldson, S. (1995). The backpropagation algorithm: Implications for the biological bases of individual differences in intelligence. *Intelligence, 21*(3), 327-345.

Anderson, J. A. (1995). *An introduction to neural networks.* Cambridge, MA: The MIT Press.

Anderson, J. A., & Rosenfeld, E. (1998). *Talking nets: An oral history of neural networks.* Cambridge, MA: The MIT Press.

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

Anderson, J. R. (1987). Production systems, learning, and tutoring. In D. Klahr & P. Langley & R. Neches (Eds.), *Production system models of learning and development* (pp. 437-458). Cambridge, MA: The MIT Press.

Anderson, J. R. (1993). *Rules of the mind.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1995). *Cognitive psychology and its implications* (4th ed.). New York: W. H. Freeman.

Anderson, O. R. (1992). Some interrelationships between constructivist models of learning and current neurobiological theory, with implications for science education. *Journal of Research in Science Teaching, 29*(10), 1037-1058.

Anderson, S. J., & Conway, M. A. (1997). Representations of autobiographical memories. In M. A. Conway (Ed.), *Cognitive models of memory. Studies in cognition* (pp. 217-246). Cambridge, MA: The MIT Press.

Atran, S. (1995). Causal constraints on categories and categorical constraints on biological reasoning across cultures. In D. Sperber & D. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 205-233). Oxford, UK: Clarendon Press/Oxford University Press.

Atran, S. (1996). From folk biology to scientific biology. In D. R. Olson & N. Torrance (Eds.), *Handbook of education and human development: New models of learning, teaching and schooling* (pp. 646-682). Malden, MA: Blackwell Publishers.

Atran, S. (2002). Modular and cultural factors in biological understanding: An experimental approach to the cognitive basis of science. In P. Carruthers & S. Stich (Eds.), *Cognitive basis of science* (pp. 41-72). Cambridge, UK: Cambridge University Press.

Averbach, E., & Sperling, G. (1961). Short term storage of information in vision. In C. Cherry (Ed.), *Information theory* (pp. 196-211). London: Butterworth.

Bailey, C. H., & Chen, M. C. (1983). Morphological basis of long-term habituation and sensitization in aplysia. *Science, 220*, 91-93.

Baker, E., Croot, K., McLeod, S., & Paul, R. (2001). Psycholinguistic models of speech development and their application to clinical practice. *Journal of Speech Language & Hearing Research, 44*(3), 685-702.

Baker, J. C., & Martin, F. G. (1998). *A neural network guide to teaching* (Phi Delta Kappa Fastback #431). Bloomington, IN: Phi Delta Kappa Educational Foundation.

Baker, M. (1994). A model for negotiation in teaching-learning dialogues. *Journal of Artificial Intelligence in Education, 5*(2), 199-254.

Banich, M. T. (1997). *Neuropsychology: The neural bases of mental function*. New York: Houghton Mifflin.

Bear, M. F., Connors, B. W., & Paradiso, M. A. (1996). *Neuroscience: Exploring the brain*. Baltimore, MD: Williams & Wilkins.

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the mind: Introduction to parallel processing in networks*. Cambridge, MA: Basil Blackwell.

Bereiter, C. (1991). Implications of connectionism for thinking about rules. *Educational Researcher, 20*(3), 10-16.

Bereiter, C., & Scardamalia, M. (1996). Rethinking learning. In D. R. Olson & N. Torrance (Eds.), *The handbook of education and human development: New models of learning, teaching, and schooling* (pp. 485-513). Malden, MA: Blackwell Publishers.

Berg, T., & Schade, U. (2000). A local connectionist account of consonant harmony in child language. *Cognitive Science, 24*(1), 123-149.

Berlyne, D. E. (1965). *Structure and direction in thinking*. New York: Wiley.

Besner, D., Twilley, L., McCann, R., & Seergobin, K. (1990). On the association between connectionism and data: Are a few words necessary? *Psychological Reports, 97*, 432-446.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. New York: Oxford University Press.

Blake, D. T., Strata, F., Churchland, A. K., & Merzenich, M. M. (2002). Neural correlates of instrumental learning in primary auditory cortex. *Proceedings of the National Academy of Sciences, 99*(15), 10114-10119.

Bollaert, M. (2000). A connectionist model of the processes involved in generating and exploring visual mental images. In S. O. Nuallain (Ed.), *Spatial cognition: Foundations and applications* (pp. 329-346). Amsterdam, Netherlands: John Benjamins Publishing Company.

Brady, J. A. (1995). *Connectionist models of reading and spelling transfer.* University of Delaware (Unpublished Thesis), Delaware, MD.

Bramer, M. A. (1982). Pattern-based representations of knowledge in the game of chess. *International Journal of Man-Machine Studies, 16*(4), 439-448.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school.* Washington, D.C.: National Academy Press.

Broadbent, D. (1985). A question of levels: Comment on McClelland and Rumelhart. *Journal of Experimental Psychology: General, 114*(2), 189-190.

Brown, J. C. (2005). When standing at a crossroads, how do we decide the right path?, *PsycCRITIQUES* (Vol. 50): American Psychological Association.

Bruer, J. T. (1993). *Schools for thought: A science of learning in the classroom.* Cambridge, MA: The MIT Press.

Bruer, J. T. (1997). Education and the brain: A bridge too far. *Educational Researcher, 26*(8), 4-16.

Bruer, J. T. (1999a). In search of... Brain-based education. *Phi Delta Kappan, 180*(9), 649-657.

Bruer, J. T. (1999b). *The myth of the first three years: A new understanding of early development and lifelong learning.* New York: Free Press.

Bruer, J. T. (2002). Avoiding the pediatrician's error: How neuroscientists can help educators (and themselves). *Nature Neuroscience Supplement, 5*(November 2002), 1031-1033.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* New Brunswick, NJ: Transaction Books.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Carver, S. M., & Klahr, D. (Eds.). (2001). *Cognition and instruction: Twenty-five years of progress.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Case, R. (1996). The role of central conceptual structures in the development of children's thought. *Monographs of the Society for Research in Child Development, 61*(1-2), 1-24.

Catterall, W. A. (1993). Structure and function of voltage-gated ion channels. *Trends in Neurosciences, 16*, 500-506.

Charness, N. (1992). The impact of chess research on cognitive science. *Psychological Research/Psychologische Forschung, 54*(1), 4-9.

Chase, W. G., & Simon, H. A. (1988). The mind's eye in chess. In A. M. Collins & E. E. Smith (Eds.), *Readings in cognitive science: A perspective from psychology and artificial intelligence* (pp. 461-494). San Mateo, CA: Morgan Kaufmann, Inc.

Chomsky, N. (1959). A review of B. F. Skinner's 'verbal behavior'. *Language, 35*(1), 26-58.

Chown, E. (2004). Cognitive modeling. In A. B. Tucker (Ed.), *Computer science handbook, 2nd edition* (Chapter 69). Boca Raton, FL: Chapman & Hall/CRC.

Chua, S. L., Chen, D.-T., & Wong, A. F. L. (1999). Computer anxiety and its correlates: A meta-analysis. *Computers in Human Behavior, 15*(5), 609-623.

Churchland, P. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy, 78*(2), 67-90.

Churchland, P. M. (1988). *Matter and consciousness*. Cambridge, MA: The MIT Press.

Churchland, P. M. (1995). *The engine of reason, the seat of the soul: A philosophical journey into the brain*: Cambridge, MA: The MIT Press.

Churchland, P. S. (1986). *Neurophilosophy: Toward a unified science of the mind/brain*. Cambridge, MA: The MIT Press.

Clark, A. (1987). From folk psychology to naive psychology. *Cognitive Science, 11*(2), 139-154.

Cohen, I. L., Sudhalter, V., Landon-Jimenez, D., & Keogh, M. (1993). A neural network approach to the classification of autism. *Journal of Autism and Developmental Disorders, 23*(3), 443-466.

Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing* (pp. 151-216). San Diego, CA: Academic Press.

Coltheart, M. (1985). Cognitive neuropsychology and the study of reading. In M. I. Posner & O. S. M. Marin (Eds.), *Attention and performance xi* (pp. 3-37). Hillsdale, NJ: Erlbaum.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel distributed processing approaches. *Psychological Review, 100*(4), 589-608.

Congdon, R., & Raudenbush, S. (2001). Opdes - optimal designs (Version 0.19): HLM Software (available at http://www.ssicentral.com/other/hlmod.htm).

Connell, M. W. (2002). *On the viability of computational neuroscience as a framework for connecting brain, mind, and education*. Unpublished qualifying paper, Cambridge, MA: Harvard University Graduate School of Education.

Cosmides, L., & Tooby, J. (1994). Beyond intuition and instinct blindness: Toward an evolutionarily rigorous cognitive science. *Cognition, 50*(1), 41-77.

Craik, F. I. M. (1986). A functional account of age differences in memory. In F. Klix & H. Hagendorf (Eds.), *Human memory and cognitive capabilities, mechanisms, and performances* (pp. 409-422). Amsterdam: Elsevier.

Crown, J. (1996). Schizophrenia and frontal lobe development: A twin and sibling study of development during young adulthood. *Dissertation Abstracts International, 56*(9-B), 5163-5418.

Czaja, S. J. (1996). Aging and the acquisition of computer skills. In W. Rogers & A. D. Fisk & N. Walker (Eds.), *Aging and skilled performance: Advances in theory and applications* (pp. 241-266). Mahwah, NJ: Erlbaum.

Dehaene, S. (1996). The organization of brain activations in number comparison: Event related potentials and the additive-factors method. *Journal of Cognitive Neuroscience, 8*(1), 47-68.

Dehaene, S. (1999). *The number sense: How the mind creates mathematics*. Oxford, UK: Oxford University Press.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science, 284*, 970-974.

Demircioglu, G., Ozmen, H., & Ayas, A. (2004). Some concepts misconceptions encountered in chemistry: A research on acid and base. *Educational Sciences: Theory & Practice, 4*(1), 77-80.

Descartes, R. (1641/1960). *Meditations on first philosophy*. New York: MacMillan.

Dunn, E. W., Wilson, T. D., & Gilbert, D. T. (2003). Location, location, location: The misprediction of satisfaction in housing lotteries. *Personality & Social Psychology Bulletin, 29*(11), 1421-1432.

Eisenhart, M., & DeHaan, R. L. (2005). Doctoral preparation of scientifically based education researchers. *Educational Researcher, 34*(4), 3-13.

Elman, J. L. (1989). Connectionist approaches to acoustic/phonetic processing. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 227-260). Cambridge, MA: The MIT Press.

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

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*(1), 71-99.

Elman, J. L. (1995). Language as a dynamical system. In R. F. Port & T. v. Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (pp. 195-226). Cambridge, MA: The MIT Press.

Elman, J. (1998). Connectionism, artificial life, and dynamical systems. In W. Bechtel & G. Graham (Eds.), *A companion to cognitive science* (pp. 488-505). Malden, MA: Blackwell Publishers.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: The MIT Press.

Ermentrout, B. (1994). Reduction of conductance-based models with slow synapses to neural nets. *Neural Computation, 6*, 679-695.

Feigenbaum, E. A., & Feldman, J. (1995). *Computers and thought.* Cambridge, MA: The MIT Press.

Fiez, J. A., Raichle, M. E., Balota, D. A., Tallal, P., & Petersen, S. E. (1996). PET activation of posterior temporal regions during auditory word presentation and verb generation. *Cerebral Cortex, 6*(1), 1-10.

Fiez, J. A., Raichle, M. E., Miezin, J. D. E., Petersen, S. E., Tallal, P., & Katz, W. F. (1995). PET studies of auditory and phonological processing: Effects of stimulus characteristics and task demands. *Journal of Cognitive Neuroscience, 7*(3), 357-375.

Fischer, K. W., & Bidell, T. R. (1998). Dynamic development of psychological structures in action and thought. In R. M. Lerner (Ed.), *Handbook of child psychology (fifth edition)* (Vol. 1, pp. 467-562). New York: John Wiley & Sons.

Fischer, K. W., & Connell, M. W. (2003). Two motivational systems that shape development: Epistemic and self-organizing. *Development and Motivation - Joint Perspectives, 1*(1), 103-123.

Fischer, K. W., & Farrar, M. J. (1987). Generalizations about generalization: How a theory of skill development explains both generality and specificity. *International Journal of Psychology, 22*, 643-677.

Flavell, J. H., Miller, P. H., & Miller, S. A. (1993). *Cognitive development* (Third ed.). Englewood Cliffs, NJ: Prentice Hall.

Fodor, J. A. (1968). *Psychological explanation.* New York: Random House.

Fodor, J. A. (1975). *The language of thought.* Cambridge, MA: Harvard University Press.

Fodor, J. A. (1987). Why there still has to be a language of thought, *Psychosemantics.* Cambridge, MA: MIT Press/Bradford Books.

Fodor, J. A. (1990). Defending the "language of thought". In W. G. Lycan (Ed.), *Mind and cognition: A reader.* (pp. 282-311). Oxford, UK: Basil Blackwell, Inc.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(1), 3-71.

Forster, K. I., & Forster, J. C. (2003). Dmdx: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments & Computers, 35*, 116-124.

Galley, W. C. (2004). Exothermic bond breaking: A persistent misconception. *Journal of Chemical Education, 81*(4), 523-525.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design patterns: Elements of reusable object-oriented software.* New York: Addison-Wesley.

Gardner, H. (1985). *The mind's new science: A history of the cognitive revolution.* New York: Basic Books.

Gardner, H. (1993). *Multiple intelligences: The theory in practice.* New York: BasicBooks.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2002). *Cognitive neuroscience: The biology of the mind* (Second ed.). New York: W. W. Norton and Company.

Gelman, S. A., & Raman, L. (2002). Folk biology as a window onto cognitive development. *Human Development, 45*(1), 61-68.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Gershenfeld, N. (1999). *The nature of mathematical modeling.* New York: Cambridge University Press.

Gilbert, D. T., Pinel, E. C., Brown, R. P., & Wilson, T. D. (2000). The illusion of external agency. *Journal of Personality & Social Psychology, 79*(5), 690-700.

Gilbert, D. T., & Wilson, T. D. (2000). Miswanting: Some problems in the forecasting of future affective states. In J. P. Forgas (Ed.), *Feeling and thinking: The role of affect in social cognition* (pp. 178-197). Cambridge, UK: Cambridge University Press.

Gleick, J. (2004). *Isaac Newton.* New York: Vintage Books.

Goldberg, A. L. (2000). Test-level, item-level, and experiential differences on computerized and paper-and-pencil versions of a practice graduate record exam (GRE). *Dissertation Abstracts International, 61*(2-A), 585-869.

Goldman, A. I. (1993). The psychology of folk psychology. *Behavioral & Brain Sciences, 16*(1), 15-29-113.

Goldstone, R. (1994). The role of similarity in categorization: Providing a groundwork. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 3-28.

Goldstone, R. (1999). Similarity. In R. A. Wilson & F. C. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 763-764). Cambridge, MA: The MIT Press.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects* (pp. 23-32). New York: Bobbs-Merrill.

Gopal, H., Kleinsmidt, J., Case, J., & Musonge, P. (2004). An investigation of tertiary students' understanding of evaporation, condensation and vapour pressure. *International Journal of Science Education, 26*(13), 1597-1620.

Goswami, U. (2004). Neuroscience and education. *British Journal of Educational Psychology, 74*, 1-14.

Graham, G. (2002). Behaviorism. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy (fall 2002 edition).* http://plato.stanford.edu/archives/fall2002/entries/behaviorism/.

Granott, N. (1998). A paradigm shift in the study of development. *Human Development, 41*(5/6), 360-365.

Griffin, S. (2004). Building number sense with number worlds: A mathematics program for young children. *Early Childhood Research Quarterly, 19*, 173-180.

Griffin, S. A., Case, R., & Siegler, R. S. (1994). Rightstart: Providing the central conceptual prerequisites for first informal learning of mathematics to students at risk for school failure. In K. McGIlly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 25-49). Cambridge, MA: MIT Press.

Gruber, H. E., & Voneche, J. J. (Eds.). (1995). *The essential Piaget.* Northvale, NJ: Jason Aronson, Inc.

Hamill, J. F. (1979). General principles of classification and nomenclature in folk biology: Two problems. *Anthropological Linguistics, 21*, 147-153.

Harman, G. (1989). Some philosophical issues in cognitive science: Qualia, intentionality, and the mind-body problem, *Foundations of cognitive science* (pp. 831-848). Cambridge, MA: The MIT Press.

Haslam, N. (2005). Dimensions of folk psychiatry. *Review of General Psychology, 9*(1), 35-47.

Hatano, G., & Inagaki, K. (1994). Young children's naive theory of biology. *Cognition, 50*(1), 171-188.

Hatfield, G. (2002). Psychology, philosophy, and cognitive science: Reflections on the history and philosophy of experimental psychology. *Mind and Language, 17*, 207-232.

Haykin, S. (1999). *Neural networks: A comprehensive foundation (2nd edition).* Upper Saddle River, NJ: Prentice Hall.

Hecht, H., & Bertamini, M. (2000). Understanding projectile acceleration. *Journal of Experimental Psychology: Human Perception & Performance, 26*(2), 730-746.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation.* Reading, MA: Addison-Wesley.

Hodgkin, A. L., & Huxley, A. F. (1939). Action potentials recorded from inside a nerve fiber. *Nature, 144*, 710-711.

Holzl, E., Kirchler, E., & Rodler, C. (2002). Hindsight bias in economic expectations: I knew all along what i want to hear. *Journal of Applied Psychology, 87*(3), 437-443.

Hubel, D. H. (1995). *Eye, brain, and vision.* New York: Scientific American Library.

Inagaki, K., & Hatano, G. (2004). Vitalistic causality in young children's naive biology. *Trends in Cognitive Sciences, 8*(8), 356-362.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic Books.

James, W. (1890). *Psychology (briefer course).* New York: Holt.

Jeffress, L. A. (1951). *Cerebral mechanisms in behavior; the Hixon symposium.* New York: Wiley.

Jennings, H. S. (1906). *Behavior of the lower organisms.* New York: Columbia University Press.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of english as a second language. *Cognitive Psychology, 21*(1), 60-99.

Jones, B. H., Hill, T. R., & Coffee, D. (1998). Computer-aided negotiation for classroom instruction: Assessing neural network potential. *Journal of Educational Technology Systems, 27*(1), 55-62.

Just, M. A., & Varma, S. (2002). A hybrid architecture for working memory: Reply to MacDonald and Christiansen (2002). *Psychological Review, 109*(1), 55-65.

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (Eds.). (2000). *Principles of neural science* (Fourth ed.). New York: McGraw-Hill.

Kashima, Y., McKintyre, A., & Clifford, P. (1998). The category of the mind: Folk psychology of belief, desire, and intention. *Asian Journal of Social Psychology, 1*(3), 289-313.

Klahr, D., & MacWhinney, B. (1998). Information processing. In W. Damon (Ed.), *Handbook of child psychology: Volume 2: Cognition, perception, and language* (pp. 631-678). New York: John Wiley & Sons, Inc.

Klahr, D., & Siegler, R. S. (1978). The representation of children's knowledge. In H. W. Reese & L. P. Lipsitt (Eds.), *Advances in child development and behavior* (Vol. 12, pp. 61-116). New York: Academic Press.

Kosslyn, S. M. (1994). *Image and brain.* Cambridge, MA: MIT Press.

Kosslyn, S. M., Chabris, C. F., Marsolek, C. J., & Koenig, O. (1992). Categorical versus coordinate spatial relations: Computational analyses and computer simulations. *Journal of Experimental Psychology: Human Perception & Performance, 18*(2), 562-577.

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago, IL: University of Chicago Press.

Laakso, A., & Cottrell, G. (2000). Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology, 13*(1), 46-76.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior: The Hixon symposium* (pp. 112-146). New York: Wiley.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge, UK: Cambridge University Press.

Leitenberg, H. (1976). *Handbook of behavior modification and behavior therapy.* Englewood Cliffs, NJ: Prentice-Hall.

Lenneberg, E. (1967). *Biological foundations of language.* New York: Wiley.

Lerner, J. S., Small, D. A., & Loewenstein, G. (2004). Heart strings and purse strings: Carryover effects of emotions on economic decisions. *Psychological Science, 15*(5), 337-341.

Levin, J. (2004). Functionalism. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy (fall 2004 edition).* http://plato.stanford.edu/archives/fall2004/entries/functionalism/.

Li, Q. (2002). Gender and computer-mediated communication: An exploration of elementary students' mathematics and science learning. *Journal of Computers in Mathematics & Science Teaching, 21*(4), 341-359.

Lust, B. (2000). Requirements for paradigm shift. *Journal of Child Language, 27*(3), 744-749.

MacDonald, C., & MacDonald, G. (Eds.). (1995). *Connectionism: Debates on psycholgical explanation.* Cambridge, MA: Blackwell.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality & Social Psychology Review, 3*(1), 23-48.

Marinova-Todd, S. H., Marshall, D. B., & Snow, C. E. (2000). Three misconceptions about age and l2 learning. *TESOL Quarterly, 34*(1), 9-34.

Marinova-Todd, S. H., Marshall, D. B., & Snow, C. E. (2001). Missing the point: A response to Hyltenstam and Abrahamsson. *TESOL Quarterly, 35*(1), 171-176.

Marr, D. (1982). *Vision.* New York: W. H. Freeman and Company.

McClelland, J. L. (1989). Parallel distributed processing: Implications for cognition and development. In R. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology.* Oxford, UK: Clarendon Press.

McClelland, J. L., & Jenkins. (1991). Nature, nurture, and connectionism: Implications for connectionist models of development. In K. v. Lehn (Ed.), *Architectures for intelligence: The twenty-second (1988) Carnegie symposium on cognition.* Hillsdale, NJ: Erlbaum.

McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2). Cambridge, MA: The MIT Press.

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(4), 636-649.

McGilly, K. (Ed.). (1995). *Classroom lessons: Integrating cognitive theory and classroom practice.* Cambridge, MA: The MIT Press.

McKnight, K. S., & Walberg, H. J. (1998). Neural network analysis of student essays. *Journal of Research and Development in Education, 32*(1), 26-31.

McLeod, P., Plunkett, K., & Rolls, E. T. (1998). *Introduction to connectionist modelling of cognitive processes.* New York: Oxford University Press.

Mead, S. E., Batsakes, P., Fisk, A. D., & Mykityshyn, A. (1999). Application of cognitive theory to training and design solutions for age-related computer use. *International Journal of Behavioral Development, 23*(3), 553-573.

Medin, D. L., & Atran, S. (2004). The native mind: Biological categorization and reasoning in development and across cultures. *Psychological Review, 111*(4), 960-983.

Merzenich, M. M., Jenkins, W. M., Johnston, P., Schreiner, C., Miller, S. L., & Tallal, P. (1996). Temporal processing deficits of language-learning impaired children ameliorated by training. *Science, 271*, 77-84.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Mitchell, D. B., Brown, A. S., & Murphy, D. R. (1990). Dissociations between procedural and episodic memory: Effects of time and aging. *Psychology and Aging, 5*, 264-276.

Morlet, T., Norman, M., Ray, B., & Berlin, C. I. (2003). Fast ForWord: Its scientific basis and treatment effects on the human efferent auditory system. In C. I. Berlin & T. G. Weyland (Eds.), *The brain and sensory plasticity: Language acquisition and hearing* (pp. 129-148). Clifton Park, NY: Delmar Learning.

Mulford, D. R., & Robinson, W. R. (2002). An inventory for alternate conceptions among first-semester general chemistry students. *Journal of Chemical Education, 79*(6), 739-744.

National Research Council. (2002). *Scientific research in education*. Washington, DC: National Academy Press.

Newell, A. (1980). Physical symbol systems. *Cognitive Science, 4*(2), 135-183.

Newell, A., & Simon, H. A. (1961). Computer simulation of human thinking. *Science, 134*, 2011-2017.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*(3), 113-126.

Nichols, S. (2004). The folk psychology of free will: Fits and starts. *Mind & Language, 19*(5), 473-502.

Nunn, R. T. (1979). Functionalist psychology: Limiting the range of physical law. *Journal of Thought, 14*, 182-186.

Ohlsson, S. (2000). Localist models are already here. *Behavioral & Brain Sciences, 23*(4), 486-487.

Oliver, A. (2004). Testing the internal consistency of the standard gamble in "success" and "failure" frames. *Social Science & Medicine, 58*(11), 2219.

Oliver, A., Johnson, M. H., Karmiloff-Smith, A., & Pennington, B. (2000). Deviations in the emergence of representations: A neuro-constructivist framework for analysing developmental disorders. *Developmental Science, 3*(1), 1-40.

O'Reilly, R. C. (1999). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences, 2*, 455-462.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience*. Cambridge, MA: The MIT Press.

Osgood, C. E. (1953). *Method and theory in experimental psychology*. New York: Oxford University Press.

Ozmen, H. (2004). Some student misconceptions in chemistry: A literature review of chemical bonding. *Journal of Science Education & Technology, 13*(2), 147-159.

Papa, F. J., Shores, J. H., & Meyer, S. (1990). Effects of pattern matching, pattern discrimination, and experience in the development of diagnostic expertise. *Academic Medicine, 65*(9 Suppl), S21-22.

Pare-Blagoev, E. J. (2005). *Connecting neuroscience and education: Learning by example, the case of Fast ForWord.* Unpublished qualifying paper, Harvard University Graduate School of Education, Cambridge, MA.

Pavlov, I. P. (1927). *Conditioned reflexes.* London: Oxford.

Perkins, K., Gupta, L., & Tammana, R. (1995). Predicting item difficulty in a reading comprehension test with an artificial neural network. *Language Testing, 12*(1), 34-53.

Pinker, S. (1997). *How the mind works.* New York: W. W. Norton and Company.

Plaut, D., McClelland, J., Seidenberg, M., & Patterson, K. (1996). Understanding normal and impaired reading: Computational principles in quasi-regular domains. *Psychological Review, 103*, 56-115.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology, 10*(5), 377-500.

Plomin, R., & Walker, S. O. (2003). Genetics and educational psychology. *British Journal of Educational Psychology, 73*(1), 3/14.

Plunkett, K., & Elman, J. L. (1997). *Exercises in rethinking innateness: A handbook for connectionist simulations.* Cambridge, MA: The MIT Press.

Posner, M. I. (Ed.). (1989). *Foundations of cognitive science.* Cambridge, MA: The MIT Press.

Putnam, H. (1975). *Mind, language, and reality.* Cambridge, UK: Cambridge University Press.

Pylyshyn, Z. W. (1986). *Computation and cognition: Toward a foundation for cognitive science*: The MIT Press.

Quinn, P. C., & Johnson, M. H. (1997). The emergence of perceptual category representations in young infants: A connectionist analysis. *Journal of Experimental Child Psychology, 66*, 236-263.

Ramsey, W. (2003). Eliminative materialism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (fall 2003 edition).* http://plato.stanford.edu/archives/fall2003/entries/materialism-eliminative/.

Raudenbush, S. W., & Liu, X.-F. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6*(4), 387-401.

Reiner, M., Slotta, J. D., Chi, M. T. H., & Resnick, L. B. (2000). Naive physics reasoning: A commitment to substance-based conceptions. *Cognition & Instruction, 18*(1), 1-34.

Rincover, A., Newsom, C. D., Lovaas, O. I., & Koegel, R. L. (1977). Some motivational properties of sensory stimulation in psychotic children. *Journal of Experimental Child Psychology, 24*(2), 312-323.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context.* New York: Oxford University Press.

Rolls, E. T., & Treves, A. (1998). *Neural networks and brain function.* New York: Oxford University Press.

Rosch, E. (1994). Is causality circular? Event structure in folk psychology, cognitive science and buddhist logic. *Journal of Consciousness Studies, 1*(1), 50-65.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*, 386-408.

Roth, W. M. (1992). Artificial neural networks for modeling knowing and learning in science. *Journal of Research in Science Teaching, 37*(1), 63-80.

Rozell, E. J., & W. L. Gardner, I. (1999). Computer-related success and failure: A longitudinal field study of the factors influencing computer-related performance. *Computers in Human Behavior, 15*(1), 1-10.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318-362). Cambridge, MA: The MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: The MIT Press.

Ryle, G. (1949). *The concept of mind.* Chicago, IL: University of Chicago Press.

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychologist, 24*(2), 113-142.

Sanford, J. (2004). Think you're a smart investor? *Canadian Business, 77*(11), 75.

Schneider, W. (1987). Connectionism: Is it a paradigm shift for psychology? *Behavior Research Methods, Instruments & Computers, 19*(2), 73-83.

Schneider, W., & Graham, D. J. (1992). Introduction to connectionist modeling in education. *Educational Psychologist, 27*(4), 513-530.

Seidenberg, M., & McClelland, J. (1989). A distributed model of word recognition and naming. *Psychological Review, 96*, 523-568.

Sejnowski, T. J., Koch, C., & Churchland, P. S. (1988). Computational neuroscience. *Science, 241*, 1299-1306.

Sellars, W. (1954). Reflections on language games. *Philosophy of Science, 21*, 204-228.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal, 27*, 379-423; 623-656.

Shiv, B., Loewenstein, G., & Bechara, A. (2005). The dark side of emotion in decision-making: When individuals with decreased emotional reactions make more advantageous decisions. *Cognitive Brain Research, 23*(1), 85-92.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8*, 481-520.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development, 46.*

Simon, H. A. (1992). What is an "explanation" of behavior? *Psychological Science, 3*(3), 150-161.

Simon, H. A. (1999). Production systems. In R. A. Wilson & F. C. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 676-678). Cambridge, MA: The MIT Press.

Skinner, B. F. (1938a). *The behavior of organisms.* New York: Appleton-Century-Crofts.

Skinner, B. F. (1938b). *The behavior of organisms: An experimental analysis.* New York: Appleton-Century.

Skinner, B. F. (1945). The operational analysis of psychological terms. *Psychological Review, 52*, 270-277.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral & Brain Sciences, 11*(1), 1-74.

Snow, C. E. (1983). Age differences in second language acquisition: Research findings and folk psychology. In K. M. Bailey & M. H. Long & S. Peck (Eds.), *Second language acquisition studies* (pp. 141-150). Rowley, MA: Newbury House.

Snow, C. E. (1992). Perspectives on second-language development: Implications for bilingual education. *Educational Researcher, 21*, 16-19.

Snow, C. E. (2002). Second language learners and understanding the brain. In A. M. Galaburda & S. M. Kosslyn (Eds.), *Languages of the brain.* (pp. 151-165). Cambridge, MA: Harvard University Press.

Snow, C. E., & Hoefnagel-Hohle, M. (1977). Age differences in the pronunciation of foreign sounds. *Language and Speech, 20*(4), 357-365.

Snow, C. E., & Hoefnagel-Hohle, M. (1978a). Age differences in second language acquisition. In E. M. Hatch (Ed.), *Second language acquisition: A book of readings* (pp. 333-344). Rowley, MA: Newbury.

Snow, C. E., & Hoefnagel-Hohle, M. (1978b). The critical period for language acquisition: Evidence from second language learning. *Child Development, 49*(4), 1114-1128.

Snow, C. E., & Hoefnagel-Hohle, M. (1979). Individual differences in second-language ability: A factor-analytic study. *Language and Speech, 22*(2), 151-162.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs, 74*(11), 1-30.

Spitzer, M. (1999). *The mind within the net: Models of learning, thinking, and acting.* Cambridge, MA: The MIT Press.

Stahl, J. R., Thomson, L. E., Leitenberg, H., & Hasazi, J. E. (1974). Establishment of praise as a conditioned reinforcer in socially unresponsive psychiatric patients. *Journal of Abnormal Psychology, 83*(5), 488-496.

Stoianov, I., Stowe, L., & Nerbonne, J. (1999). *Connectionist learning to read aloud and correlation to human data.* Paper presented at the 21st Annual Meeting of the Cognitive Science Society, Vancouver, Canada.

Stuss, D. T. (1992). Biological and psychological development of executive functions. *Brain & Cognition, 20*(1), 8-23.

Sun, R. (1996). Hybrid connectionist-symbolic modules. *AI Magazine, 17*(2), 99-103.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.* Cambridge, MA: The MIT Press.

Tallerman, M. (Ed.). (2005). *Language origins: Perspectives on evolution.* Oxford, UK: Oxford University Press.

Temple, E., Deutsch, G. K., Poldrack, R. A., Miller, S. L., Tallal, P., Merzenich, M. M., & Gabrieli, J. D. E. (2003). Neural deficits in children with dyslexia ameliorated by behavioral remediation: Evidence from functional MRI. *Proceedings of the National Academy of Sciences, 100*(5), 2860-2865.

Tepper, J. A., Powell, H. M., & Palmer-Brown, D. (2002). A corpus-based connectionist architecture for large-scale natural language parsing. *Connection Science, 14*(2), 93-114.

Triantaphyllou, E. (2000). Connectionist-symbolic integration: From unified to hybrid approaches. *IIE Transactions, 32*(3), 277-278.

Turkle, S., & Papert, S. (1991). Epistemological pluralism and the revaluation of the concrete. In I. Harel & S. Papert (Eds.), *Constructionism* (pp. 161-191). Norwood, NJ: Ablex Publishing.

Viennot, L. (1979). Spontaneous reasoning in elementary dynamics. *European Journal of Science Education, 1*(2), 205-221.

Vitzthum, R. C. (1995). *Materialism: An affirmative history and definition*. Amherst, MA: Prometheus.

Vygotsky, L. (1986). *Thought and language*. Cambridge, MA: The MIT Press.

Watson, J. B. (1913). Psychology as the behaviourist views it. *Psychological Review, 20*(2), 158-177.

Welford, A. T. (Ed.). (1980). *Reaction times*. New York: Academic Press.

Wermter, S., & Panchev, C. (2002). Hybrid preference machines based on inspiration from neuroscience. *Cognitive Systems Research, 3*(2), 255-270.

Widrow, B., & Lehr, M. A. (1990). Thirty years of adaptive neural networks: Perceptron, madaline, and backpropagation. *Proceedings of the IEEE, 78*(9), 1415-1442.

Wiesel, T. N., & Hubel, D. H. (1965). Extent of recovery from the effects of visual deprivation in kittens. Journal of Neurophysiology, 28(6), 1060-1072.

Zago, M., & Lacquaniti, F. (2005). Cognitive, perceptual and action-oriented representations of falling objects. Neuropsychologia, 43(2), 178-188.

VITA

Michael W. Connell

1987-1991	Massachusetts Institute of Technology Cambridge, Massachusetts	B.S. June 1991
1987-1993	Massachusetts Institute of Technology Cambridge, Massachusetts	B.S. June 1993
1991-1993	Teaching Assistant Massachusetts Institute of Technology Cambridge, MA	
1993-1996	Research Assistant Massachusetts Institute of Technology Cambridge, MA	
1993-1996	Massachusetts Institute of Technology Cambridge, MA	M.S. February 1996
1994-1995	Software Design Engineer Sunburst Communications, Inc. Pleasantville, New York	
1995-1997	Software Design Engineer Microsoft Corporation Redmond, Washington	
1997-1998	Harvard University Graduate School of Education Cambridge, Massachusetts	M.Ed. June 1998
1997-2005	Doctoral Candidate Harvard University Graduate School of Education Cambridge, Massachusetts	
1998-2002	Teaching Fellow Harvard University Graduate School of Education Cambridge, Massachusetts	
1998-2001	Research Training Fellow Spencer Foundation	
2000-2004	Research Associate Lexia Learning Systems, Inc. Lincoln, Massachusetts	
2002-2005	Software Design Consultant Lexia Learning Systems, Inc. Lincoln, Massachusetts	