Exploring the limited effect of inductive discovery learning: computational models and model-based analyses
van Rijn, D.H.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Exploring the Limited Effect of Inductive Discovery Learning

Computational Models and Model-Based Analyses

Hedderik van Rijn
EXPLORING THE LIMITED EFFECT OF INDUCTIVE DISCOVERY LEARNING: COMPUTATIONAL MODELS AND MODEL-BASED ANALYSES

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof.mr. P. F. van der Heijden
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit

op dinsdag 16 september 2003, te 10:00 uur.

door Dirk Hendrik van Rijn

geboren te Utrecht
Promotiecommissie:

Promotor: prof.dr. B.J. Wielinga

Co-promotores: dr. B. Bredeweg
dr. M.W. van Someren

Overige leden: prof.dr. J.A.P.J. Breuker
prof.dr. J.J. Elshout
prof.dr. T. de Jong
dr. N.A. Taatgen

Faculteit der Maatschappij- en Gedragswetenschappen.

Printed by PrintPartners Ipskamp, Enschede.
Cover design by Simone & Hedderik.

This project was supported by a grant from the Dutch National Science Foundation (NWO/SGW, grant 575-22-004).
## Contents

### Chapter 1
**Introduction**

1. **Inductive Learning** ........................................ 1  
   1.1 **Why Study Inductive Discovery Learning?** ............. 2  
2. **Studies Reported in this Thesis** .......................... 4

### Chapter 2
**Modeling Developmental Transitions on the Balance Scale Task**  

3. **Empirical Phenomena and Criteria** ....................... 9  
   3.1 **Empirical Phenomena** .................................. 9  
   3.2 **Empirical Criteria** .................................... 15  
4. **Computational Models** ................................... 15  
   4.1 **Production Rule Models** ............................ 16  
   4.2 **Decision Tree Model** ................................ 17  
   4.3 **Connectionist Models** ............................... 18  
   4.4 **Conclusions** ........................................ 20  

5. **An ACT-R Model of Balance Scale Behavior** ............ 20  
   5.1 **Key Structures and Mechanisms of the ACT-R Architecture** 21  
   5.2 **Explanation of Development in the Model** ............ 24  
   5.3 **Simulation I: Phase-by-Phase Developmental Progression** 27
Chapter 5

Discovery skills in a complex task

6.1 Optics ................................................. 103
6.2 Empirical Results with Optics ......................... 106
6.3 Assessing Discovery Skills ............................. 108
6.4 The Qualitative Model ................................. 110
6.5 Assessing the Size of Optics’ Experiment-space ... 116
   6.5.1 Static Experiment Space ........................... 116
DANKWOORD

De volgende mensen zijn zeer invloedrijk geweest bij het tot stand komen van dit proefschrift:

- Bob Wielinga als promotor, die mij keer op keer versted heeft doen staan over zijn stuurmanschap als het weer eens duidelijk was dat dit project eigenlijk vier kapiteins had,

- Bert Bredeweg en Maarten van Someren als begeleiders en mede-kapiteins,

- Casper, Frans en Pascal als “partners in crime” en hun begeleiders, Jos Beishuizen (die mij zonder het zelf te beseffen ooit voor de wetenschap heeft weten te behouden), Wouter van Joolingen, Ton de Jong en Marcel Veenman,

- Han (en Brenda en Maartje) als inspiratoren voor het hoofdstuk waar ik het troost op ben, en Niels voor alle ACT-R kennis die ik over de jaren in Groningen heb kunnen opdoen,

- Meester Boersma, Anton Pfeiffer, Jan Pos en Wietske Vonk als mensen van wie ik soms al lang, lang geleden veel heb geleerd,

- Irma, Inge, Sanne als studenten die veel van de discussies die ik met hen heb gehad in dit proefschrift zullen terugvinden,

- Dennis en Jan als perfecte 9.13 kamergenoten en natuurlijk ook Bruno, Maarten (en indirect Michi, Jara en Alexander), Saskia en alle SWI’ers, EPOS’ers en UvA’ers die mijn tijd in Amsterdam onvergetelijk hebben doen zijn.

Echter, het afronden van een proefschrift is niet alleen het wetenschappelijke einde van een periode, maar ook een soort afronding van een meer privé tijdvak. Hoewel veel van de bovenstaande personen ook buiten-werk vrienden zijn geworden en bij mijn volgende proefschrift (ha!) ook onder deze waterscheiding staan, is er een aantal personen die niets of weinig met dit proefschrift of SWI te maken hebben gehad maar daarmee misschien nog wel belangrijker zijn geweest in de afronding ervan.

- Astrid, Carlien, Dirk, Marc, Marije (ook al spreken we elkaar veel te weinig),
Onno en Suzan als door-dik-en-dun vrienden, en natuurlijk ook Floor, als een op-het-juiste-moment-feestje-geef-vriendin,

- Christina, Dennis (en daarmee zijn Beckers-pakje), Eric-Jan, Hilde en Janneke als mensen die ik leerde kennen als collega, maar die ik nu in eerste plaats als vriend beschouw,

- Adam, Daisy, Eleanor, Galia, Mark, Niels en Steffi als new Pittsburgh friends,

- Andreas, Celine, Det, Liesbeth en Sanne (en de rest) als perfecte anti-RSI en pro-ontspanning acrobaten,

- Wim, Willie, Marieke, Sofie en Rogier – en hun partners – voor het zo’n goede familie zijn dat ik tranen in mijn ogen krijg als ik denk aan hoe blij ik ben dat jullie mijn ouders, zusjes en broertje zijn. En natuurlijk ook Utie, die ik nog steeds een beetje tot die groep vind behoren.

Maar bovenal, dank aan Simone, die mij heeft leren drijven in de zee.

Hedderik,
CHAPTER 1.

INTRODUCTION

This thesis is about inductive learning. More particularly, it is about the difficulties associated with inductive learning. Research has shown that the (educational) gain of inductive learning is not as high (see for an overview De Jong & Van Joolingen, 1998) as initially claimed (Bruner, Goodnow, & Austin, 1956). The first question addressed in this thesis concerns this topic: "Why is inductive learning less effective than claimed?"

To answer this question, the related question: "How to measure the quality of the inductive learning process?" will also be addressed. This is necessary as inductive learning is difficult both for the "learners", the people submerged in inductive learning, as for the researchers studying inductive learning. For example, there is no generally approved variable that measures the quality of the inductive learning process.

The structure of this chapter is as follows. First, inductive learning is defined. Hereeto, induction, as the process that is the basis of inductive learning, is contrasted to deduction, and inductive learning is contrasted to scientific discovery learning. Second, five reasons to study inductive learning as identified by Klahr and Simon (1999) are discussed, and related to the studies reported in this thesis. Finally, the four Chapters that each discuss a different aspect of inductive learning are introduced and discussed.

INDUCTIVE LEARNING

In their influential book on induction, Holland, Holyoak, Nisbett, and Thagard (1986) define induction as "all inferential processes that take place in the face of uncertainty" (Holland et al., 1986, p.1). In other words, induction is concerned with inferring knowledge from an incomplete set of observations. The most typical way to infer knowledge is by inducing rules that hold for a complete domain, that is, also for the parts of the domain that are not directly observed. As the resulting knowledge is based on incomplete information, it is inherently uncertain because the not observed
parts of the total set of observations might falsify the inferred knowledge.

This contrasts induction with deduction where the learner formulates regularities observed in a (complete) set of data. Deduction does not produce knowledge that is semantically new. It does not go beyond the observations. In contrast, induction, in the words\(^1\) of Johnson-Laird (1993) is: “any process of thought yielding a conclusion that increases the semantic information [that was available] in its initial observations or premises” (Johnson-Laird, 1993, p.60).

Another issue related to induction is how these observations are constructed. In most induction related tasks the learner has to engage in the discovery process without explicit guidance, i.e., the learner has responsibility for the construction of the data set (Anderson, 1995).

With the emphasis on the self-directed nature of inductive learning and the focus on inducing new rules for yet unobserved data, inductive learning is closely related to scientific discovery. Often, these two terms are used interchangeably. Whereas the terms induction and inductive learning are often referred to in the context of contrasting induction with deduction (but see also Poletiek, 2001), scientific discovery is the term most often used when this task is studied for understanding the underlying processes. The “scientific” part of this term implicitly emphasizes the importance of constructing and testing hypotheses (cf. the empirical cycle of De Groot, 1969, the often cited methodology underlying scientific discovery). In contrast, the term induction refers to the focus on gathering knowledge based on an evidence-driven instead of hypothesis-driven process. As discussed in later Chapters, learners in the tasks studied in this thesis seem to be less focused on central hypotheses. Therefore, in the remainder of this thesis, the term inductive learning will be preferred over scientific discovery. However, regardless of the name, the central issue is: Why study inductive discovery learning?

**WHY STUDY INDUCTIVE DISCOVERY LEARNING?**

In an overview paper by Klahr and Simon (1999), five reasons for studying scientific discovery or inductive learning are identified:

1. **Human Value**  By studying scientific discovery we might better understand the

---

\(^1\)See Manktelow (1999) for an extensive overview of different views on induction.
mechanisms of "scientific thinking [that] has enhanced our ability to understand, predict, and control the natural forces that shape our world" (p.524). This in itself is claimed by Klahr and Simon to be a viable reason to study scientific discovery processes.

2. **(Non-)Mythology** Although reports of scientific discovery are often surrounded with mythological explanations (from an apple that falls from a tree to the rising of the level of water in a bath), Klahr and Simon claim that "normal" cognitive processes are at the basis of most scientific discoveries. One of the aims of studying scientific discovery learning is to demystify the processes involved.

3. **Pressing the limits** Inherent to the nature of discoveries, most discoveries are related to examining the limits of the then-known theory of the domain under study. However, they are also taking place at the limits of the discoverer's knowledge and/or reasoning capacities. Nevertheless, Klahr and Simon claim that the methods used to arrive at new discoveries are not fundamentally different from everyday thinking.

4. **Relation to Development** Research has shown that although children's discovery behavior is not as sophisticated as adults' behavior, children do show behavior that supports the notion of the "child-as-scientist": actively engaging in a quest for explanations. As this behavior is not formally taught to young children, studying development in general and the development of discovery processes could yield more information about the underpinnings of scientific discovery. Moreover, when more is known about the nature and use of scientific discovery, applying this knowledge in educational settings could improve learning effectiveness.

5. **Computational Support for Scientific Discovery** By identifying the important processes of scientific discovery, computer programs can be designed on the basis of these processes to aid human scientific discovery. However, most of the current approaches in this field are mainly based on mathematical methods (e.g., inspecting deviations from correlational patterns) instead of methods based on knowledge about the human discovery processes. By learning more about the human discovery processes and implementing this new knowledge,
the supporting computer programs might become even more helpful discovery assistants.

Of these five reasons, the research reported in this thesis is mainly concerned with the reasons 2, 3 and 4. As claimed by Klahr and Simon and reflected in their second and third reason, the assumption underlying the research in this thesis is that scientific discovery can be considered a relatively "normal" process, or at least one that can be carried out by humans that are not professionally trained researchers. This topic will be the main focus of Chapters 3, 4 and 5. Chapter 2 focusses on the development-related issue. Based on detailed empirical developmental data, this chapter gives an in-depth explanation for the development from imperfect knowledge to a final set of relations that accurately describes the given domain.

**STUDIES REPORTED IN THIS THESIS**

In most of the studies discussed in this thesis, inductive learning is studied in controlled laboratory experiments. In such a laboratory setting, university students or younger children are presented with a (computer) simulation of a particular domain. By means of self-directed experimentation, they have to induce the rules underlying the domain.

In Chapter 2 we analyse development on the balance scale as inductive learning. In particular we give an explanation for well-established phenomena in development on this task. This model does not involve active experimentation because few experimental studies include this form of learning. Instead, empirical data is used that was gathered by presenting different age groups with a set of experimenter-controlled experiments. In contrast to the classic method of analyzing the inductive learning processes by correlating overt behavior phenomena with the outcomes of the learning process, Chapter 2 describes a computational model of the same task as presented to the learners. Based on an analysis of the differences observed between the age groups, and the transitions from each level of performance to a next, a computational model was constructed that accounts for the important aspects of performance. By analyzing this model, for example, through the identification of the mechanisms nec-

---

2The first of these reasons is a relative or subjective criterion, that is, what intrigues one researcher, has no impact at all on another. The fifth reason is purely practical. As this thesis is involved with explaining discovery learning, the application of knowledge about the discovery process is outside the scope of this thesis.
necessary to account for observed behavior, inferences can be made about the underlying processes in human learners. We will argue that the typical phases observed in children's behavior can be explained by assuming a simple problem-solving strategy and a gradual increase in available resources and knowledge.

A similar approach is taken in Chapter 3. On the basis of a rational analysis of a discovery task presented to college students and based on an analysis of their behavior, a family of computational models is constructed that is based on the influential Scientific Discovery as Dual Search theory (Klahr & Dunbar, 1988). By comparing the differences between models and the match of these models with data of certain learners, the relative importance of different aspects of the inductive learning process is assessed. This comparison shows that learners use appropriate discovery techniques and that scope of effects discovered is largely dependent on the learner's prior knowledge and assumptions.

In both Chapter 2 and 3, the assumptions underlying computational models are tested by comparing the behavior of the models with empirical data. A methodologically inspired stance is taken in Chapter 4. To construct a psychologically plausible computational model of an inductive learning task, detailed information has to be available about the underlying process and knowledge that guides performance. However, as in almost all psychology domains, these processes and knowledge are not readily available for observation. Therefore, particular aspects of overt behavior have been used as measures for assessing which mechanisms underly inductive learning. At the same time, measures such as the number of unique learning-instances created by a learner have been used to assess the quality of a learner's inductive learning skills. However, the computational modeling effort and task analysis presented in Chapter 3 shows that such measures are not necessarily indicative of the quality of the underlying processes. Chapter 4 proposes a measure that is based on an analysis of the subprocesses involved in inductive learning. It is argued that this measure better reflects the quality of the inductive learning process than existing measures.

Analyzing the behavior of a learner is directly related to the ease of categorizing the learner's behavior. In the tasks presented in Chapter 2, 3 and 4, as in most other published inductive learning tasks, this poses no problem as each action of a learner results in a new and distinct configuration of the experiment setting. However, in tasks that resemble real-life discovery situations, the values of variables are often
expressed on a continuous scale and modified in a continuous fashion - and therefore they are not as readily compared. Chapter 5 focuses on such a more complex task, Optics. As the important variables defining this task are indeed expressed on a continuous scale, one has to devise a method to analyze the continuous behavior. Chapter 5 introduces a qualitative reasoning model to analyze this type of discovery settings. We argue that the learners' behavior can be characterized as being based on simple strategies that are mediated by both prior knowledge and the saliency of the discrepancies between prior knowledge and observed behavior. This characterization also explains why learners often conduct insufficient experiments.

Although all chapters address both questions that were put forward at the start of this introduction, Chapter 3 and 5 are mainly concerned with the question “What causes inductive learning to be difficult?” and Chapter 4 focuses on the methodological issue of how to measure inductive learning. In Chapter 5, both questions are interwoven, as the methodological aspects discussed in that Chapter clearly show the complexity of inductive learning that would otherwise not be easy to pinpoint.

In the final Chapter of this thesis, the different approaches in the Chapters 2 to 5 are summarized and interpreted in terms of a general theory of inductive learning. As put forward in the above discussed “(Non-)Mythology” and “Pressing the limits” reasons, the research presented in this thesis does support the claim that there is nothing special about inductive learning. Even the large differences between learners found in some of the tasks reported in this thesis can be easily explained with down-to-earth arguments; as these differences are largely due to differences in the learners' assumptions associated with the discovery task - not necessarily with good or bad discovery behavior. Moreover, the research reported in this thesis suggest a refinement of the claim as put forward in the influential SDDS accounts (Klahr & Dunbar, 1988) that inductive/discovery learning is mainly driven by hypotheses. That is, instead of playing a central role, hypotheses seem to play a relative minor role compared to the major role of experiment construction and interpretation.
CHAPTER 2.

MODELING DEVELOPMENTAL TRANSITIONS ON THE BALANCE SCALE TASK

Abstract

Periods of relatively stable, rule-like behavior alternated with short transition periods characterize cognitive development on reasoning tasks like the balance scale task. Each transition gives rise to an improvement in behavior, until a phase is reached in which performance is flawless or improvement is not worthwhile given the necessary extra effort. Several computational models have been developed to capture the developmental phenomena associated with the balance scale task. These models, which originate from different computational traditions, explain the main phenomena of development. Recently, empirical phenomena have been reported that these models cannot easily accommodate. We propose a computational model that is implemented in ACT-R and that is based on the evaluation of success of applied knowledge, combined with a mechanism to construct new knowledge by searching for differences between the left- and right-hand sides of presented balance scale problems. This model accounts for the main empirical phenomena as well as for the recently reported empirical phenomena such as learning without feedback.

Cognitive development in many domains is conceptualized as a progression through a series of increasingly complex and increasingly accurate task-specific stable phases. Behavior on the balance scale task, a task that is related to proportional reasoning, displays this type of developmental progression (Chlestos, De Lisi, Turner, & McGillicuddy-De Lisi, 1989; Jansen & Van der Maas, in press; Siegler, 1981). Periods of relatively stable performance ("phases"), alternate with short periods of unstable performance during which a new phase is attained. In a balance scale task, a child is asked to predict the direction of movement of a balance scale. Pegs are situated at equal distances from each other and the fulcrum. Identical weights are placed on one of the pegs at both sides of the fulcrum. The balance will tip to either side, or remain in balance, depending on the number and the positions of the weights. This task is used to investigate the strategies that children employ in solving this task, the effect of training on the children's behavior, and the transition from one strategy to the next (e.g., see Siegler, 1976, 1981; Siegler & Chen, 1998). Siegler distinguished six types of balance scale items\(^1\), which he used to categorize behavior in terms of the inferred strategies. The six types are divided into three simple types and three so-called conflict-types. The simple item types are: (a) *Balance items* with equal numbers of weights at each side and equidistant to the fulcrum; (b) *Weight items* with unequal numbers of weights at each side, equidistant from the fulcrum; and (c) *Distance items* with an equal number of weights at each side, but at different distances from the fulcrum. These item types are called simple item types as the weight and distance information do not conflict. The three conflict item types are characterized by both different numbers of weights and different distances from the fulcrum. Moreover, the answers based on each dimension conflict: if the weight dimension predicts the answer to be "the balance scale tips to the left side", the distance predicts the opposite. Depending on the differences between the weights and distances, the scale tilts to the side with the largest number of weights or to the side where the distance to the fulcrum is greatest. In (d) *Conflict Weight items*, the balance scale tilts to the side with the larger number of weights. In (e) *Conflict Distance items*, the scale tilts to the side where the distance to the fulcrum is greater. In (f) *Conflict Balance items*, the effects of the two dimensions compensate: the scale remains in balance. The behavior of children on the balance scale task has been studied extensively by comparing their

---

\(^1\)Given a five weights and five distances at each side balance scale, 68% of all possible items fall in these six items types, see Table 2. The remaining items have more weights at the side with the larger distance, and are therefore less suitable to categorize behavior. We will refer to these items as Weight/Distance items.
performance on the different item types. These studies have established repeatedly a number of empirical phenomena which will be discussed in the next section.

Identifying phenomena does not provide an account of the mechanisms underlying behavior. As argued by many authors (e.g., Boden, 1979; Anderson, 1987; Siegler, 1989), one needs to understand the mechanisms that cause the changes in behavior to obtain a causal explanation of development. In attempts to explain these mechanisms in the case of the balance scale task, several computational models have been proposed. These models, which originate from different computational traditions, explain the main phenomena of development. Recently, new empirical phenomena have been reported which these models cannot easily accommodate.

This paper consists of three parts. We first describe the main empirical phenomena associated with development on the balance scale task. Second, we discuss existing computational models of the balance scale task. Finally, we present a computational model that does account for the major empirical phenomena and we discuss its explanatory value.

EMPIRICAL PHENOMENA AND CRITERIA

In this section, we discuss the major empirical phenomena (EP) associated with the balance scale task. On the basis of these phenomena, we formulate four sets of empirical criteria. Only if a computational model satisfies these four criteria, it can be considered a model of the full range of behavior associated with the balance scale task.

EMPIRICAL PHENOMENA

**EP1: Stable Phases and Transitions** The behavior of children on the balance scale task is often classified according to their inferred use of a small number of “Rules”\(^2\). The notion of Rule use is based on the observation of homogeneous behavior on balance scale items of a particular type. This “consistency criterion” (Reese, 1989) is often cited as evidence for rule use. Within the balance scale domain, the consistency criterion is operationalized by Siegler’s (1981)

---

\(^2\)To ease presentation, the Rules empirically identified by Siegler and others, see EP2, are written with a upper case “R”, whereas ACT-R production rules, to be introduced later, are written with a lower case “r”. A single Rule is always implemented as a set of production rules.
rule assessment methodology. Using this methodology, children's responses are compared with the responses implied by the various Rules. Their behavior is classified according to the closest match between their responses and those associated with a given Rule. Jansen and Van der Maas (1997) improved the rule assessment methodology by applying latent class analysis. Latent class analysis\(^3\) is a statistical technique used to determine the number and kind of unobservable classes of behavior which give rise to observed responses on balance scale items. Using this technique, Jansen and Van der Maas (1997, in press) provided a statistically sound foundation of Siegler's earlier results (1976, 1981).

The overall picture that emerges from these studies is that children adopt progressively more powerful Rules during development, which results in increasingly better performance. Between transitions, children display stable and consistent behavior on the balance scale task. During the transition from one Rule to the next, performance is more erratic as some items of a particular type are already answered correctly, whereas others are still answered according to the earlier Rule yielding incorrect answers. The exact mechanism of transition may differ from Rule to Rule.

**EP2: The Rules** Siegler (1976, 1981) has identified four Rules that characterize children's behavior during different phases. These Rules are shown in Figure 3.1.

Rule I involves the following steps: consider the number of weights on each side: if equal, predict that the balance scale will remain level, if unequal, predict that the balance scale will tip to the side with greater weight. Like Rule I, Rule II states that the scale will tip to the side with the greater number of weights. However, if the weights are equal, the distance dimension is taken into account. Rule III states that both the weight and distance dimension are taken into account. If one side has greater weight and the other greater distance, the child guesses. (Described as “muddle through” in Siegler, 1976.) Rule IV states that both dimensions are considered and if the dimensions conflict (as in the conflict item types), the torque is computed by multiplication of the weights and distances. Note that these Rules have multiple conditional responses, and

---

\(^3\)Latent class models consist of classes that are defined by response patterns associated with the item types. Latent class models are fitted by determining the relative size of each of the classes (McCutcheon, 1987). Main advantages over the standard rule assessment method are (a) statistical methods to evaluate the fit of the Rules to the data are available, (b) it is not necessary (although possible) to specify the Rules a priori, so that new Rules can be detected, and (c) different balance scale tests with different numbers of items can be compared.
are therefore more complex than “standard” IF-THEN rules.

The unsystematic behavior before Rule I is often described as Rule 0 or pre-Rule I. The first three Rules (Rule I-Rule III) are comparable because they consist of simple steps consisting only of comparisons. Rule IV includes an additional step that involves multiplication. Rule I is completely unidimensional, as only the weights are taken into account. Rule II is neither completely unidimensional nor multidimensional, as it can be applied by switching attention from one dimension to another. Only Rule III and Rule IV require full multidimensional reasoning as both weight and distance values are considered simultaneously (Siegler, 1996). The period in which a child uses a particular Rule is referred to as a Phase (e.g., Rule I is used in Phase I).

After Siegler’s initial study, these results have been replicated in numerous other studies. Besides confirmation of the originally presented Rules, some new Rules were proposed. However, a number of these Rules are hard to identify empirically as these predict the same performance as other Rules (e.g., see Jansen & Van der Maas, in press for a latent class based analysis and categorization of Rules used by children). An exception is the Addition Rule (Ferretti, Butterfield, Cah, & Kerkman, 1985; Normandeau, Larivee, Roulin, & Longeot, 1989). The Addition Rule states that the answer is based on the addition of weight and distance values when presented with conflict items.

It is sometimes doubted whether subjects classified as Rule IV use multiplication rather than some other procedure, such as perceptual approximation or problem-answer mappings. However, in studies in which the rule assessment classification is studied by asking for verbal explanations, most subjects classified as Rule IV refer to multiplication in their explanations (Siegler, 1976, 1981). Another issue with regard to Rule IV is whether Rule IV can be learned without explicit instruction which would have serious implications for balance scale models. As advanced subjects new to the balance scale task regularly use Rule IV to solve the problems, the Rule can be learned without explicit balance scale task instruction. Of course, it is possible that they received explicit instruction in related domains. But irrespective of how the Rule is learned, a significant group of subjects master Rule IV, and so should complete balance scale models do.
Figure 3.1. Rules that categorize behavior as identified by Siegler (1981).
Modeling Developmental Transitions on the Balance Scale Task

Figure 3.2. Three transition patterns, x-axis denotes the difference in distances left and right of the fulcrum, the y-axis denotes the used Rule. Note that the actual distance difference values are not constrained, for example, the Maxwell convention can also occur at distance difference four.

**EP3: Transition Phenomena without Feedback** In a recent study, Jansen and Van der Maas (2001) describe phenomena associated with the transition from Rule I to Rule II. Jansen and Van der Maas presented a sequence of distance items to Phase I children. The weights of the first item were situated on the first peg left from the fulcrum, and on the second peg right from the fulcrum. During the first part of the sequence, the distance between the fulcrum and the weights was increased by moving the weights on the right side one peg outward per presentation. During the second part of the sequence, which started once the weights had reached the most extreme peg, the distance was decreased again by moving the weights back toward the fulcrum. This way, the difference between the left and right distance values first increased and then decreased. The choice to manipulate distance was based on research that suggests that the *availability* or *encoding* of the distance dimension influences performance on the balance scale task (Siegler & Chen, 1998; Siegler, 1976, Exp.3). As in most balance scale studies, the children were not given any feedback concerning their responses. Of the 314 children tested (age range 6 to 10 years), 27% displayed behavior that was not consistent with Rule I over the sequence of difference distance manipulations. Three interesting deviations from Rule I were found. In these cases, children showed behavior in accordance with three transition patterns. The diagrams in Figure 3.2 illustrates these patterns and the associated responses.
Four percent of the children displayed a change from Rule I to Rule II behavior and from Rule II to Rule I at exactly the same point in the sequence. This pattern, which is shown in Figure 3.2a, is consistent with the so-called “Maxwell convention” (see Jansen & Van der Maas, 2001). Three percent of the children displayed a switch at a different point in the sequence depending on whether the difference in distances was increasing or decreasing. This pattern, which is shown in Figure 3.2b, is consistent with the so-called “Delay convention”. Nine percent of the children displayed a sudden jump from Rule I to Rule II as the distance difference was increased, but did not return to Rule I as the distance difference was decreased. This so-called “sudden-jump” response pattern is shown in Figure 3.2c. Besides these three groups, eleven percent of the children displayed behavior that deviated from Rule I, but that was not classifiable. Regrettably, such detailed information is not available for other transitions; Jansen and Van der Maas (2001) only studied the transition from Rule I to Rule II.

Note that these results do not necessarily imply that manipulating the distance dimension is the only or even the most important way to trigger transitions. However, the results do show that transitions can be induced without feedback. In general, this behavior is comparable to learning without feedback which has also been posited in areas as language learning (Pinker, 1984; Taatgen & Anderson, 2002) and implicit learning (e.g., Cleeremans & McClelland, 1991).

**EP4: Torque Difference Effect** A similar but less specific effect is the “torque difference effect”. This effect refers to the finding (Ferretti et al., 1985; Ferretti & Butterfield, 1992) that within item types, items with a large difference in torque are more likely to be answered in a manner consistent with a more advanced Rule than items with a smaller torque difference. This suggests heterogeneous behavior within item types. Such heterogeneity challenges the notion of Rules and the rule assessment method as proposed by Siegler. However, reanalysis of the data (Jansen & Van der Maas, 1997) shows that the torque difference effect only occurs with items with extreme torque differences. Therefore, it is concluded that the children's behavior is homogeneous for moderate torque difference levels and that a torque difference effect is limited to large torque differences.
EMPIRICAL CRITERIA

Ideally, these behavioral phenomena should be reproduced by models of development on the balance scale. Therefore, we translate these phenomena into four sets of empirical criteria for models of balance scale behavior.

EC1. Rule-like Behavior The behavior of the model should be classifiable into Rules. This classification can be conducted by latent class analysis or by inspecting the models' implementation directly. As different levels of rule-like behavior implies transitions, the model must explain transitions from one Rule to another.

EC2. Rule sets A complete model of balance scale behavior should include the four Rules as identified by Siegler (1976), and the later identified Addition Rule (Normandeau et al., 1989). The Rules should appear in a fixed order.

EC3. Transition Patterns without Feedback The model should explain how transitions occur on series of items with increasing and decreasing distance differences in the absence of feedback. In particular, the model should give rise to the three transition patterns as presented in Figure 3.2. This criterion applies to the transition from Rule I to Rule II as detailed results concerning the transitions between the other Rules are not available.

EC4. Torque Difference Effect for Large Torque Differences A model should reproduce the Torque Difference Effect for large torque differences but not for small difference values.

COMPUTATIONAL MODELS

Several computational models have been proposed to explain development on the balance scale. We discuss models of balance scale behavior based on production rules, decision trees, and neural network models. Without entering into the discussion on the psychological validity of these architectures, the models are discussed in terms of the degree to which the empirical criteria are satisfied.

Some models directly implement the Rules as specified in Figure 3.1. Classification by latent class analysis is superfluous in this case.
Obviously, the models should have access to some form of feedback. Without feedback, none of the models would be able to show development through the phases. However, in accordance with EC3, the models should be able to explain how local developmental transitions can take place without feedback.

**Production Rule Models**

We discuss two production rule models\(^5\). The first computational model of balance scale behavior is Klahr and Siegler's (1978) production rule model. They showed that different sets of static production rules were able to capture the observed Rules. The model did not specify a transition mechanism and it therefore failed to meet the first, most important empirical criterion (EC1).

Sage and Langley (Sage & Langley, 1983; Langley, 1987) presented the first computational model that actually shows development from Rule I to Rule III. As shown in Figure 3.1, the central idea is that Rule II adds extra conditions to Rule I, as does Rule III to Rule II. Their model starts with a set of production rules that produce random answers. The model learns by discrimination: if the model produces an incorrect response to an item, a new production rule is created based on an analysis of the differences between this item and the last item which the production rule solved correctly. Performance is kept stable for some time by a strength parameter. A new production rule is created when no production rule is available with a strength level above a certain threshold. This model shows rule-like behavior and transitions (EC1) from Rule I to Rule III. The model does not acquire Rule IV, because this Rule cannot be constructed as the concept of multiplication is not available to the model. The same holds for the Addition Rule. However, incorporation of multiplication and addition mechanisms will probably result in the construction of the appropriate Rules.

The discrimination method used in this model often gives rise to production rules that result in behavior not found in empirical data. The model therefore satisfies EC2 only partially. It is unclear how this model may incorporate the non-feedback related phenomena (EC3). The model only checks whether weights and distances differ and does not use the actual values in conditions of production rules. Therefore it is not

\(^5\)An often-cited production rule model of balance scale behavior is a model in SOAR (Newell, 1990, p.465ff). However, the description of this model does not provide enough information to evaluate it.
sensitive to differences within item types and fails to satisfy EC4.

The ACT-R architecture and the model that we present below share a number of features with this production rule model. ACT-R incorporates the idea of usefulness parameters for production rules, but in a more sophisticated way. As in the models of Sage and Langley, our model also constructs new production rules. However, as discussed below, the new production rules are based on knowledge available to the ACT-R model, instead of on purely architectural mechanisms. In the Langley (1987) model, new production rules are constructed by adding conditions to existing rules. In the ACT-R model, new production rules are created that operate in cooperation with existing rules.

**Decision Tree Model.**

Schmidt and Ling (1996) presented a model based on decision tree learning. This model acquires behavior on the balance scale by constructing decision trees. At the top-level of the tree is the most predictive test, one branch below are the next most predictive tests given the value on the top level test, etcetera. For example, a decision tree for Rule II (Schmidt & Ling, 1996, p.217) starts with “Are weights and distances equal?” If both are equal, the model predicts that the balance scale will remain in “balance”, otherwise the next test is “Which side has the most weights?”. During the construction of the tree, new conditions are added until the tree reaches a pre-specified “error tolerance” threshold. The algorithm used by Schmidt and Ling uses a complete training set when constructing the tree. However, as pointed out by the authors, algorithms exists that produce similar trees by presenting training items incrementally.

The decision tree model is similar to the production rule models of Sage and Langley. A path from the top-level of the tree to an end-point can be viewed as a rule. The concepts from which the tests are constructed are binary concepts representing (a) the side with the larger number of weights, (b) the side with the larger distance and (c) a concept that represents whether the weights and distances for both sides are equal. Besides these concepts, which were also available in the above discussed models of Sage and Langley, the decision tree model also has access to concepts expressing the (d) numerical difference in weights and the (e) numerical difference in distances between both sides.
The model satisfies EC1 as it acquires trees that correspond to the Rules. Moreover, the transitions from one Rule to another are by definition sudden. As the model learns the first three Rules in the right order, it partially adheres to EC2. However, the concepts of addition and multiplication are not included in the model. Therefore, the model is not able to learn the Addition Rule and Rule IV is approximated by specific item-to-answer mappings. If multiplication were included in the model, it is unlikely that the other strategies would remain available. It is unclear whether this model can incorporate the non-feedback related phenomena (EC3). With respect to EC4, different Rules are constructed for the values of the numerical difference concepts (d and e), hereby reproducing the Torque Difference Effect. However, the Torque Difference Effect is produced consistently, that is, both for smaller and larger torque difference values.

Connectionist Models

McClelland (1989, 1995) presented a connectionist model of balance scale behavior that learns by back-propagation of feedback. The network consists of two output nodes representing the predicted answer and ten input nodes, five for each side. Only one of the input nodes per side receives activation. The distance dimension determines which of the input units is activated, and the weight dimension determines the amount of activation. The input and output nodes are connected by separate hidden layers for the distance and weight dimensions. To explain the order of the Rules (i.e., the use of the weight dimension before the distance dimension), McClelland assumes that initially weight items occur more frequently than distance items.

The first empirical criterion poses a problem for this model. Using Siegler's rule assessment methodology, McClelland claimed a successful fit to human data. However, reanalysis of the model's behavior using latent class analysis showed that the behavior of the model cannot be described by a set of distinct rules (Jansen & Van der Maas, 1997). The model therefore fails to show homogeneity within item types. Furthermore, Raijmakers, Van Koten, and Molenaar (1996) showed that the model fails to show qualitative transitions between the Rules.

With respect to EC2, the model is not able to learn stable Rule behavior as it sometimes regresses after a period of stable behavior. Moreover, Rule IV is never learned. Because the model is only able to learn by feedback-dependent back-propagation, it
fails the no-feedback criterion (EC3). With respect to EC4, the model does show the Torque Difference Effect (McClelland, 1995).

Shultz and his co-workers (Shultz & Schmidt, 1991; Shultz, Mareschal, & Schmidt, 1994; Shultz, Schmidt, Buckingham, & Mareschal, 1995) modeled balance scale behavior with cascade-correlation networks. The central tenet of cascade-correlation is the introduction of new hidden nodes into the model if the back-propagation-like learning mechanism does not reduce the error between predictions and feedback fast enough. Here we will discuss the model as presented in Shultz et al. (1995). This model has five input units for each side of the balance scale and two output units. The distance dimension determines which of the input units is activated, and the weight dimension determines the amount of activation. Initially, as in McClelland's model, weight items are over-represented in the training problems to accommodate the initial preference for weight.

According to analyses with the rule assessment methodology⁶, the behavior over time of the cascade correlation network follows the Rules. However, because the learning mechanism of the cascade correlation network is comparable to that of McClelland's model, it is likely that the transitions are better described by gradual adaptation than by sudden transitions. This impression of gradual adaptation is reinforced by the remark made by Shultz (1994, p. 81) that: "The cascade correlation model suggests that balance scale [...] transitions are soft and tentative rather than abrupt and definitive". The threshold for adding new hidden nodes plays a role in how long performance remains stable. This model therefore occupies a position between the production rule model and the back-propagation model with respect to the generation of Rules (EC1). This model learns all four Rules in the correct order (EC2). However, as no torque is calculated, the model can only approximate Rule IV behavior. Shultz et al. (1994) did not attempt to incorporate the Addition Rule in their model. Like McClelland's model, it fails the no-feedback criterion (EC3), because the model is only able to learn by feedback-dependent back-propagation. The model does show the Torque Difference Effect (EC4).

---

⁶Latent class analysis has yet to be applied to the data generated by this model.
CONCLUSIONS

Table 3.1 summarizes the empirical criteria that each model satisfies. Although each model provides an interesting perspective on the process of balance scale development, none is able to model all empirical data. The most prominent problem is that none of the models can explain learning without feedback. In the next section we present an account of behavior on the balance scale task that satisfies all four empirical criteria. This model is based on the ACT-R cognitive architecture.

<table>
<thead>
<tr>
<th>Empirical Criteria</th>
<th>Production Rules</th>
<th>Decision Trees</th>
<th>Neural Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC1: Rule-like behavior</td>
<td>KS'78</td>
<td>SL'87</td>
<td>SL'96</td>
</tr>
<tr>
<td>EC2: Rule sets</td>
<td>✓</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>EC3: Transition Patterns without Feedback</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EC4: Torque Difference Effect</td>
<td>-</td>
<td>+</td>
<td>✓</td>
</tr>
</tbody>
</table>

Note. ✓: Criterion fully satisfied, +: Criterion partly satisfied, -: Criterion not satisfied. Note that the evaluations are based on the descriptions of the models as available in literature. This does not mean that models with "+" and "-" evaluations cannot be extended to match the criteria better.


AN ACT-R MODEL OF BALANCE SCALE BEHAVIOR

ACT-R (Anderson, 1990; Anderson & Lebiere, 1998) is a hybrid cognitive architecture in which the use of symbolic knowledge is mediated by associated quantitative values\(^7\). In constructing an ACT-R model of balance scale behavior, we distinguish between three aspects of the model: mechanisms, task-specific concepts, and capacity constraints. Most important are the mechanisms. The mechanisms involve several ACT-R specific processes, which will be explained in detail below. However, the key idea is a very general strategy, which underpins balance scale behavior in all phases. This strategy is: "solve balance scale items by searching for differences between the

\(^7\)ACT-R has previously been applied to cognitive development. Taatgen (1999) focussed on the compilation and usage of different production rules, Lebiere and Anderson (1998) focussed on the acquisition and tuning of declarative knowledge, and Jones, Ritter, and Wood (2000) focussed on capacity related constraints.
left and right side of the balance scale”. The performance of this strategy is constrained by the presence or absence of task-specific concepts and by the available capacity. With respect to the task-specific concepts, the availability of a particular concept or property determines whether it can be used in the reasoning process. Examples are the facts that multiplication is not available initially, and that children notice the weights earlier than the distances. The capacity constraint, expressed as universe versus multidimensional reasoning, influences the number of concepts (i.e., weight and distance) that can be incorporated simultaneously in the general strategy. Initially, children can only accommodate a difference on one dimension, which constrains the general strategy to “search for a difference between the left and right side of the balance scale”. We will demonstrate how the interplay between these three aspects of the model explains the four empirical phenomena. First, we give an overview of the properties of the ACT-R architecture that are important in the present model. Then, we discuss the task-specific concepts and the capacity constraints. Finally, we present simulations of the developmental process including all phase transitions, the transition patterns and the torque difference effect.

**Key Structures and Mechanisms of the ACT-R Architecture**

**Chunks & Production Rules** ACT-R stores knowledge in two types of memory. Declarative memory contains chunks that represent descriptive knowledge, whereas procedural memory contains production rules that represent procedural knowledge in the form of IF-THEN actions.

Each chunk is of a specific type that defines its function in the model. At each point in time, one of the chunks in the model denotes the goal of current processing.

At each processing step, ACT-R examines all production rules to test which production rules match the current goal. A production rule is considered to match if the conditions of that production are satisfied by the active goal. Besides conditions on the active goal, a production rule can also have conditions on the availability of declarative information. To retrieve information from declarative memory, a production rule initiates a retrieval request. This request specifies the conditions that determine the type of information that is required. Subsequent production rules can test for the availability of that knowledge, that is, they can test whether the given conditions returned a retrievable chunk. In addition to initiating a retrieval, the THEN-side of
production rules can change the current focus to a new goal, modify the current goal or initiate other external actions.

**Activation & Utility** Instances of both declarative and procedural knowledge have associated quantitative parameters that express their usefulness. If a chunk is retrieved or a production rule has proven to be successful in achieving the goal, the associated parameter is increased in value. As a consequence, the chunk or production rule will be used more often. For declarative memory chunks, this quantitative parameter is represented as the activation level of the chunk. The activation level depends on a base-level activation and on activation sources originating in the model’s context. The base-level activation \(B\) is determined by the number and recency of past retrievals of that chunk, i.e., \(B = \ln(\sum_{j=1}^{n}t_j^{-d})\). The summation is over all previous retrievals \(n\) of this chunk, with \(t_j\) being the time in seconds between retrieval \(j\) and present, and \(d\) being the decay rate (by default fixed at .5). The default ACT-R context activation, called associative activation does not play an important role in the model presented here. However, in the second set of simulations presented in this paper, a different type of context activation is introduced that reflects perceptual saliency. When a model initiates a retrieval request, and multiple chunks with activation levels above the retrieval threshold satisfy the conditions of that retrieval request, the chunk with the highest activation is retrieved.

For production rules, the main parameter is the utility. The utility of production rules is determined by a combination of the proportion of successful completions of goals and the costs associated with that production rule: \(\text{Utility} = U = P \times G - C + \sigma\). In this equation, \(P\) is the proportion of successful applications of the production rule. This proportion is multiplied by a constant \(G\) (20, by default). \(C\) reflects the cost associated with the execution of the production rule, expressed in units of time. The last variable in the equation, \(\sigma\), represents normally distributed noise that may optionally be added to the utility.

The value of the parameters \(P\) and \(C\) depend on the performance of the current production rule and the performance of the production rules that follow the current one. With respect to the \(P\) parameter, a production rule is considered to be executed successfully if the goal to which it was applied is tagged as successfully solved. The value of the \(C\) parameter depends on both the costs of the current production rule
and the costs of following production rules used for the same goal. Each production rule has a fixed cost of 50ms. Besides these fixed costs, a production rule might take more time if it has to wait for a retrieval from declarative memory, or if it involves perceptual or motor actions. During the run of a model, the P and C parameters are updated on the basis of new experiences. For example, if a large number of goals are solved successfully, P will increase as the overall proportion of correctly answered problems increases. Similarly, if the costs associated with solving a problem decreases (e.g., because of acquiring a more efficient strategy), the C parameter will decrease.

As with chunks, if more than one production rule satisfies the current constraints, the production rule with the highest utility value is selected. In the simulations reported below, normally distributed noise was added to the utility. This results in the model occasionally trying alternative production rules. If an alternative production rule performs better than the original production rule, this is reflected in its utility parameters. Eventually, this alternative production rule will supersede the original.

Learning & Production Composition  ACT-R models are able to expand their knowledge by adding new chunks and production rules to their memories. The source of new declarative knowledge can be internal or external. The internal source of new chunks are production rules. New chunks are added if the production rule modifies the goal-chunk to which it was matched, or if it constructs a new chunk in its THEN-side. An example of an external source of new chunks is perception.

A recent addition\(^8\) to ACT-R (ACT-R v5.0, production rule learning, 2002; Taatgen & Anderson, 2002; Taatgen, 2000) is production composition. Production composition creates new production rules by joining two production rules that occur in succession into one single rule. This process eliminates intermediate retrievals from declarative memory. For example, assume that two production rules are executed in succession. If the first production rule initiates a retrieval request and the second production rule processes the retrieved chunk by storing information from that chunk in the current goal, production composition constructs a new production rule that modifies the goal chunk in the same fashion as before, but without requiring a retrieval. In other

\(^8\)This addition is not part of the ACT-R version as described in Anderson and Lebiere (1998), but is incorporated in a new version (ACT-R v5.0, 2002). Although the version of ACT-R as described in Anderson and Lebiere (1998) is also able to construct new production rules, the new composition mechanism constructs new rules using a more principled method.
words, production composition specializes a general procedure to solve a problem into a problem specific procedure by eliminating the retrieval.

This composition mechanism also applies to learning new behavior from declarative descriptions of the involved actions. These representations (declarative actions) are available as chunks (cf., Anderson, 2000), for example as the results of observing others perform actions, or as the result of explicit instruction. The declarative actions can be matched and executed by "interpretive production rules". These production rules retrieve the declarative actions and modify the goal based on the contents of the declarative actions. Because of these modifications, other production rules now match the goal, giving rise to new behavior. By means of this process of alternating execution of interpreting and "normal" production rules, new sequences of behavior emerge on the basis of already available production rules.

The utilities of newly composed production rules are inherited from the parent rules. However, as the composed production rule often removed the necessity of a retrieval and combined two production rules into one, the costs associated with the new production rule are lower than the old combination of production rules. That is, the costs of the newly composed production rules does not involve the costs associated with extra retrievals (a variable amount of time) and the costs of executing the second production rule (i.e., a fixed 50 ms). As the success rate of the newly composed production rule is equal to the success rate of the non-composed production rules, the difference in costs will eventually favor the composed production rules.

EXPLANATION OF DEVELOPMENT IN THE MODEL.

Figure 3.3 shows three main factors underlying the behavior and development of the ACT-R model. The first factor, the mechanisms, is a combination of the ACT-R architecture and task-general knowledge. ACT-R provides the basis for behavior (e.g., the execution of production rules, the selection of chunks) and development (e.g., composition of new production rules, updating of the quantitative parameters that express the utility of production rules). The task-independent part of the mechanisms contains the interpretive production rules and the declarative representations of actions. These actions represent the strategy "answer a balance scale problem by

9For links to an extended discussion of composition and learning from instruction, see the ACT-R v5.0 website: http://act.psy.cmu.edu/ACT-R_5.0
searching for differences”.

However, this does not specify the property to which the model should attend. This is specified by the second factor: the task-specific concepts. The availability of a concept is mediated by the activation of the declarative chunk representing that concept. If this chunk is above the retrieval threshold, the related concept can be incorporated in the decision process. We make two, rather uncontroversial, assumptions about the second factor. The first is based on the empirical observation (Siegler, 1976; Siegler & Chen, 1998) that children initially prefer weight above distance. We simply assume an initially higher activation for weight than distance on the basis of the encoding studies of Siegler (1976). The second assumption states that the concepts of addition and multiplication appear later in development. The exact timing of the appearance is of less importance since the emergence of the Addition Rule and Rule IV also depend on the capacity constraints.

\[^{10}\text{Note that the ACT-R model would also work given an initially higher activation for distances, i.e., it does not explain the preference of weight above distance.}\]
Capacity constraints form the third factor. In many developmental theories, notably the Neo-Piagetian theories, the increase in capacity plays a central role in the development of domain specific strategies (Case, 1985; Pascual-Leone, 1970). Although it has proven to be difficult to quantify the precise effects of cognitive capacity (Chapman, 1990), it is quite clear that its increase plays an important role in the transition from uni- to multidimensional thinking (Siegler, 1996). It is this specific consequence of the capacity constraint that we use to explain the generalization of the “search for differences” strategy. Initially, capacity limitations constrain the strategy to “search for a difference”, and only later is the strategy extended to “search for more than one difference”.

The development of performance depends on these three factors. The mechanisms explain how development occurs, the other two factors explain the order and timing of development.

From its initialization, the model has access to declarative knowledge about how “forced choice problems” are to be solved. This declarative knowledge represents the strategy that one has to search for a property that differs for the answer alternatives (e.g., “tip to left”, “tip to right” or “remain in balance”) to solve a force-choice problem. In the context of the balance scale this means that it should look for a difference between the left- and right-hand side of the balance scale. If a difference is found based on the interpretation of the declarative actions, the model uses this difference to determine an answer. However, if no difference is found, the model searches for a new property.

Initially, neither the weight nor the distance property is encoded because the activation levels of the chunks representing the property are below threshold. Therefore, the model generates an answer by guessing. After each guess, feedback is given about the correctness of the given answer. As the proportion of correct answers based on guessing is low, the production rules representing this strategy have a low utility. Therefore, as soon as the weight property becomes available, the model will start to incorporate this concept in the decision process, yielding Rule I behavior. As this increases the proportion of correct answers, the production rules associated with Rule I are preferred over the pre-Rule I production rules. As the proportion of successful answers is still relatively low, distance is incorporated shortly after it becomes available to the model. However, as the available capacity is insufficient to
incorporate both weight and distance in the decision process at the same time, the model switches its attention to distance, discarding the weight information. As determining the answers based solely on weights is less successful if the weights are equal than if the weights are unequal, the shift from weights to distances only occurs if the weights are equal. When the capacity has increased sufficiently to make it possible to use a strategy that incorporates both weights and distances, the model progresses to Rule III behavior. In this Phase, the weights and distances are examined regardless of the weights being equal or unequal. However, as knowledge to combine the weights and distances is as yet unavailable, the model can only guess the answer if the weights and distances are both unequal. Only when the concept of addition or multiplication becomes available is the model able to progress to the Addition Rule or Rule IV.

**Simulation I: Phase-by-Phase Developmental Progression**

In this section we present the model in more detail and discuss how it simulates development. In discussing the model, we will refer to Table 3.2 and Figure 3.4. Table 3.2 presents the numbers and proportions of correctly answered problems for all possible balance scale configurations. The numbers presented in this Table are specific to a balance scale with a maximum of five weights and five distances on either arm. For Rule I and Rule II, Table 3.2 contains both the overall statistics and statistics broken down into the two different branches of these Rules (i.e., “weights equal” and “weights unequal”, see Figure 3.1). As all items of a particular item type are answered using the same branch, the other branch does not apply for that item type. This is denoted by a full stop in Table 3.2.

Figure 3.4 plots the utilities of key production rules as function of the number of balance scale problems presented to the model. As the random selection of training items influences the exact course of the utilities, the graph shows some variation over runs. However, the values shown are illustrative for a typical run of the model. The production rules in Figure 3.4 are all constructed by the model on the basis of the declarative actions.

**Phase 0:** At the start of each presentation, a problem is randomly sampled from the set of all possible problems. For each problem, the model is presented the number of weights and the distance to the fulcrum for both sides. Initially, the model resorts
Table 3.2: Number of correct responses per combination of Rule and balance scale item type for a five weights times five distances balance scale configuration.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Balance Scale Types</th>
<th>Overall Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>D</td>
</tr>
<tr>
<td>pre-Rule I (guess)</td>
<td>8.3*</td>
<td>33.3*</td>
</tr>
<tr>
<td>Rule I</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>weights equala</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>Rule II</td>
<td>25</td>
<td>100</td>
</tr>
<tr>
<td>weights equala</td>
<td>25</td>
<td>100</td>
</tr>
<tr>
<td>Rule IIIb</td>
<td>25</td>
<td>100</td>
</tr>
<tr>
<td>Addition Rulec</td>
<td>25</td>
<td>100</td>
</tr>
<tr>
<td>Rule IV</td>
<td>25</td>
<td>100</td>
</tr>
<tr>
<td>Number of items</td>
<td>25</td>
<td>100</td>
</tr>
</tbody>
</table>

Note. Different balance scale configurations lead to different absolute and relative values. B=balance item, D=distance item, W=weight item, CB=conflict balance item, CD=conflict distance item, CW=conflict weight item, WD=weight/distance item. Divisions by three indicate guessing the answer.

*Assuming a 1/3 probability of a correct guess.

a125 items have equal weights, 500 items have unequal weights.

bAs described in Siegler (1976), e.g., “Muddle Through”

cAs described in Normandeau et al. (1989);

to guessing because it cannot retrieve any properties (i.e., the activation level of the weight property is still below threshold) to use in the problem solving process. As can be seen in Table 3.2, the proportion of correct answers for the guess-strategy is low. Therefore, the utility of the production rule that implements the guessing (see Figure 3.4, labeled A(g)) drops rapidly. Because of the low utility of this guess-rule, the interpretive production rules will often be selected. Consequently, as soon as a new strategy to solve the problems becomes available, the model will solve balance scale problems using the new strategy.

Phases I:

The general strategy is represented in five steps:

11Given the general nature of the interpretive production rules, the usage of these production rules is not limited to the balance scale domain. Because of this extensive “external” usage, the parameters will not be influenced significantly by the application of the production rules in the balance scale domain. Therefore, we assigned a constant utility of approximately 13.3 (+ noise) to these production rules. This value is based on a P value of 2/3 and a relative low C value, but other values would not influence the general behavior of the model.
1. Retrieve a property on which the two sides of the balance can differ.
2. Encode the values of the specified property.
3. If the encoded values are unequal, then base the answer on the observed difference.
4. If the encoded values are equal, then search for a new property and if found, return to Step 2.
5. If no new property can be found, then render an answer based on the encoded values.

This general strategy first leads to Rule I. Step 1 determines the property on which the model will focus its attention. As initially the weight property is most active, the model will encode the weight values (Step 2) and use those to solve the balance scale problem. If the perceived values differ, Step 3 applies and an answer is determined on the basis of this difference (predicting a tilt to the side with the greater number of weights). If the perceived values are equal, the interpretation of the declarative knowledge leads to an attempt to retrieve a new property (Step 4). However, because the activation level of the distance property is too low, the model cannot retrieve it and consequently applies Step 5. As a result of the weights being equal, Step 5 renders “balance” as answer.

When the corresponding declarative actions are interpreted, the composition mechanism constructs new production rules. As discussed earlier, the composition process removes most retrievals from the constructed production rules, resulting in hard-coded values in the new production rules. Because of this, Step 1 is replaced by a production rule that immediately initiates a retrieval of the weight values, removing the requirement of having to retrieve a property first (same holds for Step 3). As the encoding of the values is based on perception instead of on a retrieval, this step is not removed by the composition mechanism. Because in this phase Step 4 always results in a failure to retrieve a new property, the composition mechanism will remove the retrieval request for a different property and, will instead directly answer “balance” (i.e., use Step 5) if the weights are equal.

The resulting production rules$^{12}$ are presented below. Before the model reaches the

$^{12}$The production rules presented here have been edited for reasons of readability. All code necessary to run the model is available at http://www.van-rijs.org/hedderik/bstm/.
stage in which it only uses these production rules, it will have constructed numerous production rules that were only used once before being replaced by a new production rule. These replaced production rules remain available to the model. However, because these production rules (and the initial interpretive production rules) require more steps and more retrievals before arriving at an answer, these production rules are more expensive and therefore have a lower utility. As a result, the production rules associated with Rule I, as shown below, will prevail.

**Production Rule: Initialize** (\(\rightarrow W\))
IF the goal is to solve a balance scale problem
THEN request the encoding of the values of the weight property

**Production Rule: Initiate-Encoding**
IF the goal contains a request to encode the values of a property
THEN encode those values

**Production Rule: Test-Encoded-Values-a** (\(<>W:A(W)\))
IF the values of the weight property have been encoded
AND the values are unequal
THEN render as response the side with the largest value

**Production Rule: Test-Encoded-Values-b** (\(=W:A(b)\))
IF the values of a property have been encoded
AND the values are equal
THEN render as response “balance”

Each production rule has been assigned a shorthand notation which is mentioned directly after the name of the production rule (e.g., “\(\rightarrow W\)”). This shorthand notation is used both in the text and in Figures 3.4 and 3.5. In this notation, “\(W\)” stands for weights, “\(D\)” for distances, “\(g\)” for guess, “\(b\)” for balance and “\(m\)” for multiplication. The operator “\(\rightarrow\)” refers to “switch attention to”, “\(=\)” to a condition requiring equal values for the property mentioned after the equal sign, “\(<>\)” to requiring unequal
values and "A(x)" refers to "answer based on x". For example, "<>W,<>D:A(g)" refers to the production rule reflecting the following knowledge: if the weight and distance values are unequal, guess an answer.

The first production rule (->W) causes the model to consider the weight property. The second production rule initiates the encoding of the values of the weight property. These two steps are not merged because the perception based encoding step is not removed by the composition mechanism. Based on the encoded values, one out of two production rules matches. If the weights are equal, the production rule =W:A(b) renders "balance" as response, if the weights are unequal, the production rule <>W:A(W) renders as response the side with the largest number of weights.

As can be seen in Table 3.2, if the answer is based on Rule I "weights equal" (which corresponds to the =W:A(b) production rule), the proportion of correct answers is lower than if the answer is based on Rule I "weights unequal" (<>W:A(W)). Therefore, the utility of the weights equal production rule (=W:A(b)) decreases faster than the utility of the weights unequal production rule (<>W:A(W)), as is visible in Figure 3.4.

As the utility of a constructed production rule is based on the utilities of its parent production rules, newly composed production rules have different initial utilities. As the ->W production rule is based on production rules used in Phase 0, it starts with a low utility, whereas <>W:A(W) and =W:A(b) are not directly based on production rules with decreased utilities. Therefore, the utilities of these production rules are initially higher.

**Phase II:** Because of the decrease in utility of the =W:A(b) production rule, the interpretive production rules will eventually try to construct new strategies based on the declarative "search for a difference" action. The number of problems required for this to happen depends on the distribution of the balance problems. For example, if the proportion of weight items is large, as in the simulations of McClelland (McClelland, 1995, 1989), the production rules used in Rule I will encounter fewer failures and will remain active longer.

To find a new strategy, the model has to include a new property in the reasoning process. As long as the activation of the distance property is below the retrieval threshold, the model is in Phase I. Only when the distance property reaches a suffi-
Figure 3.4. Typical course of utility changes over time of selected production rules. The content of the production rules is represented in the labels, see the text for explanation.

Note. The labels of the production rules "->W" and "<>W->D" are repeated twice.

cient level of activation, is it incorporated in the answering strategy\(^{13}\). As soon as this level is reached, the composition mechanism starts to construct new production rules as it did at the start of Phase I. The resulting two most significant production rules are \(=W,<>D:A(D)\) and \(=W,=D:A(b)\). The first production rule determines the solution based on the distance values if the weights are equal. The second production rule renders the solution "remains in balance" if the values for both properties are equal. Since these production rules always yield the correct answer (see Table 3.2, Rule II, weights equal), the utility of these production rules increases asymptotically as the P value in \(U=P \times G \cdot C\) approaches 1 and the costs are constant. This increase in P also influences the utility of the \(->W\) production rule, as this rule now gives rise to a greater proportion of correct answers.

\(^{13}\)A discussion of how this threshold is reached is outside the scope of this paper, but both developmental changes in activation or threshold parameters, and exposure to tasks in which distance is an important predictor might play a role.
Phase III: During Rule II behavior, the utility of $<>W:A(W)$ decreases. This decrease will cause the interpretive production rules to be used occasionally. However, the available declarative knowledge in Phases I and II limits the “search for a difference” strategy to one unequal dimension. Therefore, as long as more advanced versions of this knowledge are unavailable, the model cannot progress to Rule III.

As the progression from Rule II to Rule III is associated with a shift from unidimensional to multidimensional reasoning (Case, 1985; Chapman, 1990; Pascual-Leone, 1970; Siegler, 1996), the declarative actions are modified to reflect this change. At about problem presentation 1250 in Figure 3.4, the declarative representations were extended to allow for the retrieval of a new property even if the values of the first encoded property were unequal. This extension involved the modification of the original Step 3, as is shown below:

1. Retrieve a property on which the two sides of the balance can differ.
2. Encode the values of the specified property.
3a. If the first encoded values are unequal, then store the encoded values and search for a new property and return to Step 2.
3b. If both the first set of encoded values and the second set are unequal then search for and apply a method that uses both sets to render an answer.
4. If the encoded values are equal, then search for a new property and if found, return to Step 2.
5. If no new property can be found, then render an answer based on the encoded values.

Now, in case the weights are unequal, applying the interpretive production rules will lead the model to apply Step 3a. This results in the retrieval of distance as an additional property to take into account. If the encoded distance values are equal, the model will base the answer solely on the weight values as these are already known to be unequal. However, if the distance values are unequal, the model searches for a method to combine both weights and distances. If a method cannot be found, the model will resort to guessing. When the composition process is complete, i.e., when no new production rules can be created, the model solves balance scale problems...
using the production rules as shown in Figure 3.5.

The guessing production rule is incorrect in two thirds of all cases. As a consequence, its utility decreases to a relatively low level. Therefore, this phase is relatively unstable: if too many incorrect answers are registered in succession, behavior will temporarily regress to Rule II (c.f., Jansen & Van der Maas, in press, Table 7). In Figure 3.4, this is reflected in the continuation of updating of the utility of the $<\rightarrow W:A(W)$ production rule. This decrease in utilities also causes the model to regularly test for other methods to combine the unequal values of weights and distances.
Phase IV: As soon as the concepts of addition or multiplication become available, these are incorporated in the decision process (Step 3b), thereby replacing the guess production rules. In Figure 3.4, the multiplication method \((<>W,<>D:A(m))\) is used from about problem 1400 onward. As a side effect, it is only after the multiplication rule has become available, that the \(<>W->D\) production rule wins the utility competition over \(<>W:A(W)\). This is visible in Figure 3.4 by comparing the utilities of \(<>W:A(W)\) and \(<>W->D\). During Phase III, their utilities are relatively equal, only after the transition to Rule IV, the utility of \(<>W->D\) increases sharply. A similar effect is visible for \(->W\) (label: \(->W'\)). As this production rule is never involved in any incorrect answer in Phase IV, its utility increases consistently. However, because of the relative large number of incorrect answers \(->W\) was involved in previously, its utility increases more gradually.

As can be seen in Table 3.2, the Addition Rule gives a correct answer to most balance scale problems. Therefore, its utility will not decrease below the utility level of the interpretive production rules. Only if the model is presented a sequence of items that are answered incorrectly using the Addition Rule, the utility of this production rule drops below the utility of the interpretive production rules. Only in this situation, the model can progress to Rule IV. This is consistent with the infrequency of a spontaneous improvement from the Addition Rule to Rule IV.

Note that even when the multiplication method is available, problems that can be solved with simpler comparisons of values will still be solved by searching for differences. This is due to the high utility of the associated production rules. Therefore, the interpretive production rules cannot interfere because the interpretive production rule's utility is too low compared to the utility of the comparison production rules. Although a different explanation of the sparse use of the multiplication method could refer to the relative high cost of multiplication, this model does not depend on this assumption.

Simulation II: Transitions without Feedback and the Torque Difference Effect

As was shown by Jansen and Van der Maas (in press), transitions from Rule I to Rule II may also take place without feedback. They describe three different types of transition patterns (see Figure 3.2). The Maxwell convention pattern obtains when...
the improvement to Rule II occurs at the same distance difference as the regression to Rule I. The delay convention pattern obtains when the improvement to Rule II occurs at a higher distance difference than the regression to Rule I. Finally, the sudden jump pattern obtains when after the progress to Rule II, the child never regresses to Rule I. In the present section, we will illustrate how transitions without feedback as well as the observed transition patterns can be explained by the model.

We presented the model with the same sequence of problems as in the experiments of Jansen and Van der Maas. Like Jansen and Van der Maas, we provided no feedback about the correctness of the given answers. Without feedback, the utilities of production rules do not change. As the activation of chunks is based solely on the number of retrievals, irrespective of the success or failure of the retrieving production rules, the activations are updated. Therefore, the only source of changes during the presentations of problems without feedback is the activation of declarative knowledge.

The manipulated variable in the experiments of Jansen and Van der Maas is the difference in distances from the fulcrum between the left and right sides of the balance scale. A number of consecutive balance problems with equal weight configurations, but with different distance configurations, were presented to children. In the first half of the problems, the difference in distance between left and right was increased by one distance unit per problem. In the second half, the distance was decreased by one distance unit per problem. We assumed that this step-by-step increase in difference would increase the perceptual saliency of that property. This additional saliency can be represented in the ACT-R model as an increased activation of the distance property. To achieve this, the activation formula (Activation = B = \ln(\sum_{j=1}^{n}t_j^d)) was extended to: Activation = B + saliency. The saliency is calculated by simply scaling the distance difference (saliency = \Delta \text{distance}/\epsilon, where \epsilon is the scaling constant). Note that the default ACT-R activation formula contains a term reflecting noise. This simulation is run without noise as noise would not provide different results in this simulation.

To progress to Rule II, the model has to apply the interpretive production rules at least occasionally. As long as the distance property remains below the retrieval threshold, performance is consistent with Rule I. However, the saliency associated with extreme distance differences can increase the activation of the distance property to a level above the retrieval threshold. If the distance becomes sufficiently activated,
the distance property is retrieved and the distance values are used in determining a solution. If this happens before the maximum distance difference is reached, the next problem is also solved using the distance property, as the previous presentation increased the base level of the distance chunk's activation (B) and the increased distance difference also increased the activation. When the maximum difference is reached, the next presented problem has a decreased distance difference and therefore a lower additional saliency activation. What happens at this point depends on the interplay between increase in base level activation and the decrease in saliency activation. If the base level was increased more than the decrease in saliency activation, the model is still able to retrieve the distance property. This results in continued Rule II based behavior. However, if the increase in B is smaller than the current decrease in saliency activation, the model is not able to retrieve the distance property, and therefore regresses to Rule I.

![Distances](https://via.placeholder.com/150)

*Figure 3.6. Simulation of no-feedback transition patterns.*

The results of the simulation are presented in Figure 3.6. The solid lines represent the activation of the distance chunk. When the total activation of this chunk is above the threshold which is depicted by the horizontal dotted line at Y=0, performance is in accordance with Rule II. The performance, given the current activation, is shown as small circles, either at the bottom of the graph (denoting Rule I use), or at the top of the graph (denoting Rule II use).

If during Rule II use the increase in the base level is small compared to the dif-
ferences in saliency activation, the regression to Rule I behavior takes place at the same distance difference as the change to Rule II behavior. The phenomenon that the change takes place at the same point is referenced to as the Maxwell convention (see Figure 3.6a). If the increase in the base level is large, the activation without the saliency component might be large enough to ensure the activation is above the retrieval threshold (the sudden jump pattern, Figure 3.6c). If the base level activation is raised slightly, but not sufficiently to be retrieved in the absence of additional saliency activation, the delay convention pattern appears (Figure 3.6b). These differences in activation effects can be explained by, for example, differences in original base-levels or by differences in levels of elaboration of declarative chunks.

Although these explanations of the transition patterns are formulated in terms of the activation of the distance property, the composition process can also explain the sudden jump pattern. During the increased activation period, the composition process starts to construct new production rules. If the composition process is complete, the resulting production rules are completely specialized. That is, all retrievals (except for the encoding of the actual values of weights and distances) are removed from these production rules. Because the explicit search for the distance property is removed from the production rules, the activation of the distance property chunk does not determine the behavior of the model. Thus giving rise to a sudden jump.

The saliency of the distance property, modeled as an additional activation of the corresponding chunk, also explains the torque difference effect. As torque difference is highly correlated with distance difference, items with large torque differences will on average cause an increased activation for the distance chunk. This increases the probability that distance will be included when the declarative actions initiate a new search for a difference. However, this will only apply if the utility of the interpretive production rules and the utility of the rules that currently solve the problem are similar. In other words, the torque difference only influences the model's behavior in the instable period just before transitions. This more restricted torque difference effect is in accordance with the reanalysis of the torque difference effect (Jansen & Van der Maas, 1997).

In this second set of simulations, the explanation of developmental phenomena is based on the influence of the additional saliency based activation. In the first set of simulations, saliency activation was disabled. However, no qualitative change in
model performance would occur if saliency was added to the model used in the first set of simulations. The effect would be limited to minor modifications in the timing of Rule transitions. As this would complicate matters unnecessarily, we decided not to incorporate the notion of saliency in the first set of simulations.

**DISCUSSION**

In this paper we presented a model of balance scale behavior that is based on three components: architectural and task-independent mechanisms, task-specific concepts, and constraints related to cognitive capacity. Earlier, we discussed the merits of previous models of balance scale behavior in terms of four empirical criteria. We now turn to a discussion of the new model in terms of the empirical criteria and subsequently we outline the relation between our model and the previous models.

**Empirical Criteria**

**EC1: Rule-like Behavior** The first empirical criterion is the presence of Rules (i.e., stable performance) and transitions between these Rules. As demonstrated in Simulation 1, the ACT-R model shows stable performance when the composition process has constructed a fixed set of production rules to solve balance scale problems. As long as the utility of the production rules is higher than that of the interpretive production rules, no new production rules are constructed. But even if the interpretive production rules are applied, no new rules can be constructed in the absence of new knowledge for the construction of better rules. In both situations, behavior is stable. However, as soon as the utility of the composed production rules drops below the utility of the interpretive production rules and sufficient knowledge is available, new production rules are composed. This changes the way in which the model responds. At this point several rules have comparable utilities: the old production rules, the interpretive production rules, and the newly composed production rules. This makes behavior at this point highly dependent on factors like noise, the distribution and type of presented balance scale problems, and the visual saliency of task dimensions.

In short, the Rule-like behavior is explained by architectural mechanisms: The stable periods associated with Rule-like behavior are explained by the stabilizing effects of the adjustments of the utility and activation of available knowledge as specified by ACT-R. Whereas the transitions to new Rules depend on the ACT-R mechanisms to
construct new production rules. With different task-specific knowledge, the framework of the model can be applied to different reasoning tasks that are associated with Rule-like behavior.

EC2: Rule Sets  A model of balance scale behavior should reproduce the four distinct Rules as originally presented by Siegler (1976), and the later identified Addition Rule (e.g., Normandea et al., 1989). As discussed earlier, our model reproduces the Rules identified by Siegler and is also able to perform according to the Addition Rule. The Addition Rule is not easily replaced by Rule IV because the Addition Rule results in the correct responses to most items. The model progresses to Rule IV only if it is presented with a carefully selected set of balance scale problems that are answered incorrectly with the Addition Rule. This is consistent with the observation that very few children progress to Rule IV spontaneously.

The order of transitions matches the empirical observed sequence. This is based on constraining the availability of task-specific concepts and the assumption of processing limitations related to cognitive capacity. For example, only when the weight concept is available, is the model able to progress from pre-Rule I to Rule I; only when the concept of multiplication becomes available, can the model progress to Rule IV, etc. With respect to the cognitive capacity, as long as the available capacity constrains the model to the use of a single property (i.e., either weight or distance), the model cannot proceed to Rule III.

As discussed earlier, the assumptions with respect to the initial unavailability of task-specific concepts and the capacity constraints are in line with empirical evidence.

EC3: Transition Patterns without Feedback  Transitions from Rule I to Rule II can occur when children are presented with a sequence of distance problems with increasing distance differences, even in the absence of any feedback about the correctness of their responses. As we have shown in Simulation 2, the ACT-R model explains this phenomenon by incorporating visual saliency in the activation of chunks. The larger the difference between the values of a property for the left and right side of the balance scale, the larger the additional activation of the chunk that represents that property. When the distance difference is decreased again, the activation of a chunk might drop below the retrieval threshold, making that property unavailable to
the answering process. Given individual differences in both activation updating and production rule composition, the three types of transition patterns emerge (i.e., the Maxwell convention, the delay pattern, and the sudden jump pattern).

However, feedback still plays an important role in the presented model. That is, only by the feedback-driven updating of the production rules utilities, the model can develop to a situation in which a transition without feedback can take place. Although feedback is not necessary for the actual transitions, a model would not progress from the beginning of a phase to the end of that phase without feedback.

**EC4: Torque Difference Effect for Large Torque Differences** As the transition patterns are directly related to the distance difference and therefore to the torque difference, our model shows torque difference effects. However, our model does predict that the torque difference effect will not be strong enough to give rise to an improvement at all times. That is, the torque difference effect can only influence the behavior of the model in the vicinity of transitions.

Summarizing, we have presented a model that passes the four empirical phenomena associated with balance scale learning. In the construction and description of the model, care was taken to clearly state all necessary assumptions and to have all assumptions supported by empirical evidence.

**COMPARISONS AND PREDICTIONS**

The successful reproduction of the empirical phenomena by the presented model was partly realized with features that were also used in previous models. Like the symbolic models, behavior in the ACT-R model is based on the application of production rules, which results in rule-like behavior. As in the previous models, new production rules are learned by extending the already present knowledge. However, instead of containing a few complex production rules (like the complete trees presented in Figure 3.1), our model consists of a larger number of smaller production rules. Each of these production rules performs only a small part of the complete answering process. Therefore, newly constructed production rules can simply replace older production rules instead of requiring a complex mechanism to modify existing production rules.
As in the neural net models of balance scale behavior, quantitative information plays an important role in the ACT-R model (i.e., utility and activation). The dynamics of these quantitative variables are important for the description of the empirical phenomena.

The combination of features from ACT-R and the symbolic and neural net type of models provides the basis for a number of achievements specific for this model: (1) The model produces the relatively abrupt transitions, which are problematic in the non-symbolic models. (2) It explains transitions without feedback and the related transition patterns, which cannot be explained from the learning methods used in the previous models. (3) The model demonstrates that, given a non-biased training set, Rule IV will not easily be learned because of the high success-rate of the Addition Rule. (4) The presented model is able to explain both phenomena related to long-term development and phenomena which are only observable during short time-spans (c.f., Anderson, 2002). (5) The model makes explicit that the notion of "search for differences" combined with a gradual increase in capacity and knowledge is sufficient to explain development on the balance scale task.

However, the most distinct feature of our model is the parsimony of its main assumption: Children who are solving balance scale problems are explicitly looking for differences between the left and right side of the balance scale.

The model also makes novel predictions about behavior on the balance scale task:

During pre-Rule III behavior, the model is not sensitive to (multiplication) instruction on the balance scale task. Only when the model’s behavior is in accordance with Rule III, the model is able to benefit from multiplication instruction. Similar reasoning holds during other Phases, for example, the model predicts that instruction emphasizing the importance of distance during stable Rule I usage does not have an effect on behavior.

The model predicts an important role for visual saliency, especially during transitions. An example of which is the explanation of the Torque Difference Effect as this effects are explained on the basis of visual saliency. Because effects of visual saliency can only influence the behavior of the model during the less stable periods around
transitions, the model predicts that the Torque Difference Effect is limited to these periods.

Although not discussed in this paper, the model can also be used to predict reaction time patterns. As ACT-R specifies the amount of time necessary for the steps in the answering process, reaction time patterns can be easily derived from the model. For example, during and just after transitions, the model predicts increased reaction times as the model is still in the process of constructing the optimal set of production rules for the new phase. During the process of constructing this optimal set, additional production rules are used, causing an increase in reaction times. Moreover, previous work on arithmetic skills (e.g., Lebiere & Anderson, 1998) might be incorporated into the model to use analyses of reaction times to shed light on what strategies are used during the unstable Phase III.
CHAPTER 3.

MODELS OF INDUCTIVE LEARNING

Abstract

In this chapter, a family of computational models of inductive learning is presented that is based on the influential SDDS theory of discovery learning. Based on a combined analysis of computational models and exemplary think-aloud data, it is concluded that the prominent role SDDS reserves for hypothesis formation is not necessarily reflected in all discovery learning tasks. Learners in this chapter’s task are more focussed on the construction of correct experiments and deriving knowledge from these experiments than on guiding their behavior on a central hypothesis as is suggested by the SDDS theory.

Inductive learning has been the focus of numerous studies in psychology (see De Jong & Van Joolingen, 1998, for an extensive overview). In this type of research, focus has often been on identifying and contrasting discovery strategies with respect to the optimality or sub-optimality of their associated outcomes. A measure is necessary that distinguishes between good and bad discovery behavior to relate the outcomes of the inductive learning process to the strategies used. As we argue in this chapter (and, from a different point of view in Chapter 4), and illustrate by analyzing computational models of inductive learning, the study of the inductive learning process

This chapter has greatly benefited from the work of and discussions with Sanne Nolst Trenité. Parts of this and the next chapter are based on Wilhelm, Beishuizen, and Van Rijn (in press) and Hulshof, Wilhelm, Beishuizen, and Van Rijn (in press).
is often hindered by the lack of appropriate descriptive measures of the quality of the inductive learning process. More specifically, we will argue that there is too much emphasis on completeness (i.e., "how much is learned") as measure to assess the quality of the discovery skills. By analyzing the subtasks of the discovery process, we will propose a measure that more accurately reflects the learner's inductive learning quality than using a single figure that supposedly accounts for the influence of all inductive learning subtasks.

This chapter is closely related Chapter 4. In the current chapter, we discuss the influential Scientific Discovery as Dual Search (Klahr and Dunbar, 1988, SDDS) description of scientific discovery and present a family of computational models inspired by SDDS. Our goal is to identify the factors that determine the behavior and performance in inductive learning tasks by means of an analysis of SDDS and the application of its underlying principles in computational models. In Chapter 4, we will discuss the empirical results of the studies conducted with a relatively simple inductive learning task and discuss why the currently used measures are inappropriate.

In this chapter, SDDS will be applied to an inductive learning task and will be used it to identify the factors that determine how much is discovered. Hereto, we analyze inductive learning behavior of learners engaged in the Peter-task (Wilhelm, 2001), described later.

**THE SDDS FRAMEWORK**

A well known description of the inductive discovery process is the SDDS framework of Klahr and Dunbar (1988), based on work by Simon and Lea (1974), and later extended by Schunn and Klahr (1995) and Van Joolingen and De Jong (1997). This theory assumes that the behavior of inductive learners is hypothesis driven (c.f., De Groot, 1969). All these hypotheses are represented in a hypothesis-space. When evidence becomes available for or against a hypothesis, this hypotheses is accepted, modified or rejected. Therefore, searching for a correct set of relations to describe the domain can be viewed as a search through this hypothesis space. The experiments conducted to test the hypotheses are represented in an experiment space. By conducting experiments, learners cover parts of the experiment space. Inferential processes extract information from the covered experiment space, yielding new knowledge.
All versions of the SDDS theory are based on the structure depicted in Figure 4.1 which can be seen as a task-decomposition of the Inductive Learning process. This representation is supposed to be read from left-to-right, depth-first. Three major parts are distinguished by Klahr and Dunbar (1988): (1) finding and specifying a hypothesis ("Search Hypothesis Space"), (2) constructing and conducting experiments ("Test Hypothesis") and (3) deriving knowledge from the experiments and modifying the hypothesis accordingly ("Evaluate Evidence"). A fourth, implied but underspecified part of the theory as described in Klahr and Dunbar (1988) is (4) determining when to stop experimenting. A reason why this fourth part is underspecified might be that in the BigTrak studies, the original task to which SDDS was applied, the decision about whether or not to stop is relatively straightforward as hypotheses are easily falsified. As soon as learners in a BigTrak task are able to correctly predict the effect of a given button in a (small) number of experiments, they have effectively solved the problem. Probably because of the simplicity of the stop decision, this part of the inductive learning process did not get a lot of attention in the SDDS theory. However, as discussed in this chapter and Chapter 4, deciding when to stop does play an important role in inductive learning.
Before we will discuss these SDDS parts in detail, we will give a walk-through of Figure 4.1 and that way describe the SDDS process. According to the SDDS theory, hypotheses are pivotal to the discovery process. Each discovery is necessarily initiated by having a hypotheses. This is the reason that **search hypothesis space** is the first step in Figure 4.1. A hypothesis consists of a frame, describing the general structure of the hypothesis (e.g., a linear main effect) and a set of slot values (which variables are the dependent and independent variables). There are two methods to **generate a frame**. Given the domain of the task, learners might already have prior knowledge about which frames might be appropriate, in which case a frame can be **evoked**. If such knowledge is lacking, the learners have to **induce a frame** by constructing “outcomes” (i.e., experiments) and generalizing over these outcomes. To **generate an outcome**, the learner has to do an **e-space move**, in other words, construct a new experiment. The first step is to **focus** on the current experimental state and **choose** an variable to change, and then actually select and **set** a value for that variable. Only after that has been done is a learner able to **run** the experiment and **observe** its outcomes. Only after observing several of these generated outcomes, the learner can **generalize these outcomes** to a new frame. Given this frame, SDDS can **assign slot values**, either based on **prior knowledge** or by using experimental **outcomes**. Either **old experimental outcomes** can be used to assign slot values, or new experimental **outcomes can be generated** in a similar way as during the generation of a frame. If a complete hypothesis has been created, the next step is to **test that hypothesis**. This is done by first selecting an **e-space move** (like earlier described), then **making a prediction** and **running** that experiment and observing **its outcomes** and matching those outcomes to the hypothesis. Given that this outcome is evidence pro or contra the hypothesis, the next step is to **evaluate the evidence**. Hereto the **outcomes are reviewed** in light of the hypothesis, and based on this review a **decision** is make to either accept, reject or modify the hypothesis. Figure 4.1 does not show the stop criterion, but after having evaluated the evidence, a decision has to be made whether to continue experimenting.

The four parts of inductive learning are discussed in more detail below.

**Hypothesis formation** In the original work of Klahr and Dunbar (1988), the first step of each scientific discovery process is defined as “find a hypothesis”. This hypothesis is the corner-stone of all processes until a new hypothesis is selected.
The initial hypothesis can be based on prior knowledge about the domain or, if such information is not available, experiments are conducted to provide information for constructing the initial hypothesis. If the hypothesis that is to be constructed is not the initial hypothesis, the conclusions derived from previous experiment can also be used as guides for the construction of the new hypothesis.

To reformulate this part of SDDS: a set of methods \((M_1)\) produces a hypothesis \((H_i)\) based on the prior knowledge \((PK)\) and the conclusions \((O)\) derived from previous experiments.

**Experiment construction** The second subtask is to construct and conduct experiments to shed light on the hypothesis constructed earlier. Because SDDS is proposed as a domain neutral theory, it does not specify in detail how experiments can or should be constructed. SDDS assumes that a learner tests whether or not the experiment has already been conducted, as only new experiments are considered to provide new and informative information. Note that this both assumes a stable world (e.g., no learning effects) and one in which the outcome of a given experiment is completely deterministic (e.g., no random component in the observed outcomes).

This part of SDDS can be reformulated into: a set of methods \((M_2)\) produces a new experiment \((E_i)\) based on the hypothesis \((H)\), prior knowledge \((PK)\) and already conducted experiments \((E)\).

**Knowledge derivation** The third subtask is to derive knowledge ("conclusions", \(C\)) from the newly gathered experiments, and to adjust the current hypothesis if necessary. This process is obviously guided by the experiments conducted by the learner. Moreover, the learners’ hypothesis determines how the experiments are interpreted. But also prior knowledge and earlier conclusions might influence the process as they might lead to certain types of confirmation bias. For example, if an earlier conclusion was to accept a certain hypothesis, the learner might be so strongly attached to this conclusion that new evidence contradicting the hypothesis is ignored.

Therefore, this part of SDDS can be reformulated as: a set of methods \((M_3)\) yields a conclusion \((C_i)\) based on the hypothesis \((H)\), the prior knowledge \((PK)\),
the conducted experiments \( E \) and the prior conclusions \( C \).

**Stop decision** The fourth subtask is concerned with the decision when to stop experimenting. Obviously, a learner should not stop experimenting if the current hypothesis is not yet rejected or accepted, nor should one stop if the probability of other possible hypotheses is evaluated as being relatively high. This last aspect can be either related to the prior knowledge a learner has, or be based on the conclusions derived from earlier experiments combined with prior knowledge.

This leads to the following formation this part of the inductive learning process: a set of methods \( M_4 \) determines whether to stop experimenting based on the status of the hypotheses \( H \) and on an evaluation of prior knowledge \( PK \) and the conclusions from conducted experiments \( C \).

Although these four steps might seem to be directly related to the schematic overview of SDDS presented in Figure 4.1, their relation is less straightforward than it might seem. First and foremost, Figure 1 is both an overview of the SDDS theory and a flowchart of behavior. Learners are assumed to first search the hypothesis space to come up with a hypothesis (which might render conducting new experiments necessary), then to test the hypothesis by conducting experiments, and then review the outcomes and modify, accept or reject the hypothesis. This data-flow approach makes the modularity of the process less opaque. For example, both the generation of the hypothesis and the testing of that hypothesis involves conducting experiments. Therefore, a more modular approach removes redundancy from the description as shown in Figure 1. A second difference between Figure 1 and the above description of the four steps, and especially their rendering as shown in Table 1 is that Figure 1 does not make a distinction between the methods and the knowledge used to perform the different parts of the SDDS process.

The SDDS model was developed in the context of the BigTrak task. This task deviated somewhat from other inductive learning tasks as it is focused on a single relatively abstract rule. This emphasizes the need of constructing relative complex hypotheses, as by “just doing experiments” a learner is unlikely to solve the task because of two reasons. First, as naive learners often have a hypothesis space that only partly overlaps with the target model (Van Joolingen & De Jong, 1997), they often
have problems finding a hypothesis that is consistent with a given set of observations. Second, as BigTrak experiments generate data that is often not easily inspected or related to the actual inputs, it is often difficult to test the current hypothesis.

The relative emphasis on hypothesis testing is illustrated by the computational implementation of parts of the SDDS theory as reported in Schunn and Anderson (1999). This implementation focuses on issues of data interpretation in the context of hypothesis-driven scholarly scientific experimentation. However, many discovery tasks are characterized by simpler observations and simple hypotheses that explain a large part of the observations. Such a task is the Peter-task, which is the focus of this study and is described in the next section.

THE PETER TASK

The task that was chosen to study inductive learning is the Peter-task. The properties of this task remove a number of the complexities encountered when studying inductive learning in a more complex setting (c.f., Hulshof, 2001; Prins, 2002, and see also Chapter 5 of this thesis). Nevertheless, research has shown this task to be a viable task for studying inductive learning in different experimental settings, with different age-groups and using different domains (Wilhelm, 2001). In this task, the learner is told a cover-story about a boy, Peter, biking to school. The learner's task is to discover what determines Peter's arrival time at school. This task is modeled after the studies conducted by Kuhn, Garcia-Mila, Zohar, and Andersen (1995). In these studies, learners are presented five variables that can be manipulated to test the effect of different levels of these independent variables\(^1\) on a dependent variable. In the studies discussed in this chapter, the learners were presented a similar task using a computerized authoring environment (FILE, implemented by Jan Wielemaker, see Hulshof et al., in press, for a description). The data discussed in this chapter were collected in two different studies, one conducted at the Leiden University by Wilhelm et al. (in press) and one conducted at the University of Amsterdam (see for a description of that experiment Schoutsen, 1999). The interface as presented to the

\(^1\)In this thesis, basic statistical terminology, similar to what is common in the description of ANOVAs, will be adopted to describe the task. This results in using the terms dependent and independent variables, describing the variables in terms of the different levels that they can take, and express effects as either a main-effect or a \textit{nth-order} interaction. However, note that in an ANOVA, an effect is expressed in terms of differences in averages associated with levels, whereas in this thesis the effects are expressed compared to a base-level.
learners is depicted in Figure 4.2.

In the Peter-task, five possible causes (i.e., variables) for Peter's being late at school are presented to the learners, without stating whether they have an effect on the time he needs to get to school. First, his choice of bicycle, racing bike or normal bike. Second, how he ate his breakfast (at home or during biking). Third, the number of books he takes with him (all or just the necessary books for that day). Fourth, the type of shoes he wears (normal shoes or sports shoes). And fifth, a three level variable, the type of bag he takes with him (book bag, backpack or sports bag). Note that in other studies (Niewold, 1998; Hulshof, 2001) with the Peter task, the three level variable was concerned with the speed at which Peter biked. One of the levels of this variable was “biking with his friends”. From the analysis of think aloud protocols, it was noted that learners were surprised and distracted by the finding that the choices made by Peter also influenced the biking-time of his friends. Therefore, this variable was changed into a variable that only influenced Peter, but had similar effects as the replaced variable.

These five dependent variables make up for a total of 48 unique experiments (i.e., \(2^4 \times 3\)). After selecting a level for each of the variables, the learners are asked to predict the outcome after which the number of minutes Peter took to bike to school is presented to the learners.

The learners were told that they had to discover how the choices related to Peter’s biking time by constructing “rows of choices”. The learners were also informed that there were five possible durations it took Peter to bike to school: 35, 40, 45, 50 and 55 minutes

Figure 4.2 shows the interface of the Peter task after a learner has constructed three experiments, and is half-way constructing the fourth. The interface consists of a number of pictures that depict the levels of each variable and an area in which both the experiments are constructed and the results are shown. Of each constructed experiment, the five selected levels are shown, together with two times in the answer area to the right of the five levels. The larger number in the center of the answer area is the time it took Peter to bike to school. (Note that the results appear immediately,

\[\text{Note that in the other Peter-tasks, the travel time was expressed as the minutes Peter arrived too late at school. It turned out that this encouraged learners to approach the task in an engineering approach ("I'm trying to get Peter to arrive in time") instead of trying to discover the effect of the variables.}\]
there is no (scaled) real-time effect in determining the time necessary for biking.) The smaller number in the lower-right corner is the predicted time (entered by the learner before the answer is shown). After constructing the fourth experiment, the top-most experiment scrolls off the screen. Previous experiments can be brought back in view by using the scrollbar at the right of the experiment window.

Unknown to the learners, only three of the five variables have an effect on the outcome. The default biking time is 35 minutes. The bag-variable has a curvilinear effect; the book bag has an additional effect of 5 minutes, whereas the other two levels do not have an effect. The bike and breakfast variables form a first order interaction. The normal bike has an additional effect of 10 minutes, but the breakfast variable only has an effect of 10 additional minutes if the racing bike level is chosen. An overview of these effects is shown in Table 4.1. For example, the score of 50 in Experiment 2 in Figure 4.2 is based on a base biking time of 35 minutes, increased with 15 minutes because Peter eats his breakfast on his racing bike, thereby negating
Table 4.1: Effects of variables in Peter-task. All combinations of levels not shown in this table do not modify the time it takes Peter to bike to school.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Levels</th>
<th>Restricting condition</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bike</td>
<td>Racing bike</td>
<td>Breakfast on bike</td>
<td>+15</td>
</tr>
<tr>
<td></td>
<td>Normal bike</td>
<td></td>
<td>+10</td>
</tr>
<tr>
<td>Breakfast</td>
<td>On bike</td>
<td>Racing bike</td>
<td>+15</td>
</tr>
<tr>
<td>Books</td>
<td></td>
<td></td>
<td>none</td>
</tr>
<tr>
<td>Shoes</td>
<td></td>
<td></td>
<td>none</td>
</tr>
<tr>
<td>Bag</td>
<td>Book bag</td>
<td></td>
<td>+5</td>
</tr>
</tbody>
</table>

Its advantage over the normal bike.

As is discussed in (Wilhelm et al., in press) and (Hulshof et al., in press), this task setup can easily be used to different domains. Besides the above described “Peter biking to school” task, a second, isomorph “Peter goes shopping” task was presented to learners. In this task, learners had to discover what the contribution of different products was on the overall price that had to be paid when buying these products. Like in the Peter biking task, the underlying model contained one interaction (the whole wheat and pear are cheaper when bought together) and one main effect (the Coca-Cola was more expensive than the Fanta or the Sprite).

For an overview of the quantitative experimental results, see (Wilhelm et al., in press) or (Schoutsen, 1999). In the remaining part of this chapter, we guide the implementation of computational models by referring to think-aloud protocols of learners working in the “Peter biking to school” and the “Peter goes shopping” tasks.

**Computational Models of Inductive Learning**

In this section, we will present a family of computational models, based on the underlying principles from SDDS. The SDDS framework provides a general outline of the inductive learning process, emphasizing the importance of searches in the experiment and hypothesis space. According to Klahr and Dunbar (1988), the discovery process is mainly guided by the (form of the) chosen hypothesis. In this chapter, we will argue that learners’ behavior is not necessarily focused on one central hypothesis, but that the structure of the task determines how the task is approached. In the context of the Peter-task, this results in a process of sequentially examining the effect of the five variables. That is, instead of an iterative process that focuses on
a central hypothesis that is tested, gets revised and finally accepted, learners in the Peter-task are initially focused on discovering the effects of the five variables individually, which does not require extensive hypothesis construction or testing. Only after testing the “main effects” of a variable, some learners test for higher level effects (i.e., interactions).

Although the SDDS framework specifies a general structure of the inductive discovery process, it is more an outline of the process than a complete specification. For example, it does not specify how experiments are constructed or on what basis hypotheses get formed. Therefore, the four subtasks of the SDDS process need to be filled in for the particular task.

Below, we will describe on the basis of fragments from think-aloud protocols how each of the subtasks are performed by learners in the context of the Peter-task. The computational models discussed later in this chapter are based on these descriptions. Table 4.2 summarizes the input and output of these subtasks as earlier discussed, as these will be important for the current discussion,

<table>
<thead>
<tr>
<th>Hypothesis formation</th>
<th>$M_1(PK + C) \rightarrow H_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constructing an experiment</td>
<td>$M_2(PK + E + H) \rightarrow E_i$</td>
</tr>
<tr>
<td>Deriving knowledge from experiments</td>
<td>$M_3(PK + E + H + C) \rightarrow C_i$</td>
</tr>
<tr>
<td>Determining when to stop</td>
<td>$M_4(PK + E + H + C) \rightarrow \text{stop}$</td>
</tr>
</tbody>
</table>

*Note: See text for explanation of the abbreviations.*

**Hypothesis Formation**

Figure 4.1 gives an impression of the complexity of the hypothesis formation (“Search Hypothesis Space”) in the original SDDS model. However, the high level of complexity as shown in this Figure is not necessary for simpler tasks like the Peter-task. For example, learners in the Peter-task did not seem to have difficulties generating the right frame for their hypotheses.

Indicated by the summary of this subtask, $M_1(PK + C) \rightarrow H_i$, the methods that construct a new hypothesis or revise the current hypothesis ($H_i$) can utilize prior knowledge ($PK$) and earlier inferred conclusions ($C$). As the Peter-domain is one with which the learners have extensive experience in their own lives, this experience
(i.e., prior knowledge) plays an important role during hypothesis formation. For example, the learners are likely to have knowledge available that can be used to construct an initial hypothesis for the variables in this task.

An example of a learner using prior knowledge to guide the formation of a hypothesis is shown below. In this excerpt\(^3\), the learner is looking for a configuration in which it takes Peter less time to arrive at school. The learner states that a likely candidate for taking less time would be a different level for the variable breakfast. After conducting the experiment, the learner is surprised that the hypothesis that having breakfast at home did not take less time, as was expected on the basis of prior knowledge.

<table>
<thead>
<tr>
<th>Learner 3</th>
<th>Experiment 2:</th>
<th>Outcome:</th>
<th>Pred.:</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Preset]</td>
<td>![Initial Hypothesis]</td>
<td>50</td>
<td>55</td>
</tr>
</tbody>
</table>

**Protocol:**

Nou bijna goed, 50. Ehm, nu is er nog een ding wat minder tijd kost en ik denk dat dat misschien wel [...] het ontbijt [is]. Nou laat ik eerst het ontbijt maar proberen, doe ik het ontbijt anders. De rest hetzelfde. En verwacht ik dat hij er 55 minuten over doet.

Well, almost correct, 50. Euhm, now there is one thing that costs less time, and I think that it might be [...] breakfast. Well, let's try breakfast first, I'll change breakfast. All the others the same. And then I expect that he'll need 55 minutes.

<table>
<thead>
<tr>
<th>Experiment 3:</th>
<th>Outcome:</th>
<th>Pred.:</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Preset]</td>
<td>![Initial Hypothesis]</td>
<td>50</td>
</tr>
</tbody>
</table>

**Protocol:**

En weer 50. Das raar, (stilte) want de rest heb ik hetzelfde gelaten. 50 again. That's odd (silence), because I did not change anything else.

With respect to the conclusions (C) from earlier experiments and hypotheses, learners tend to assume that a variable has no effect if they did not discover a main effect for that variable. The C component to \( M_1 \) is mainly of a constraining nature. However, C might also influence the construction of new hypotheses if the outcomes of

---

\(^3\)Both Dutch and English translation are presented in the think-aloud excerpts.
experiments and the hypothesis are not in line with each other. In the excerpt shown below, the learner is trying to find the fastest biking time. When it is shown that Peter’s biking time is not minimal after selecting the levels of which the learner thinks that they represent the fastest options, the learner starts to reason about hypotheses with interactions between the levels of different variables.

<table>
<thead>
<tr>
<th>Learner 2</th>
<th>Experiment 2:</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Bike" /> <img src="image" alt="Egg" /> <img src="image" alt="Box" /> <img src="image" alt="Nut" /></td>
<td>Outcome: Pred.:</td>
</tr>
<tr>
<td><img src="image" alt="Learner" /></td>
<td>50 35</td>
</tr>
</tbody>
</table>

Protocol:
Nou dat blijkt niet de snelste methode, dus een aantal dingen moet ik veranderen waardoor hij sneller zal zijn. Nou ik blijf bij de racefiets, dat lijkt me het meest logisch, ehm, dan laat ik hem nog steeds zijn boterhammen meenemen, of op de fiets eet hij ze, dan laat ik hem ze thuis opeten, want onder het fietsen kost dat alleen maar tijd. Ehm, ik laat hem nog een keer zo min mogelijk boeken mee nemen, en nog steeds de sportschoenen aandoen, en z’n rugzak meenemen want dan hoop ik dat eten van invloed is.

Well, it seems that that is not the fastest method, so I’ll have to change a number of things to make him go faster. I stick with the racingbike, that seems to be most sensible, ehm, I’ll stick to have him eat his bread on the bike, oh, he eats them on his bike, then I will have him eat them at home, because eating on his bike will only cost time. Ehm, I’ll have him take as few books with him as possible, and still wear his sporting shoes, and have him take his backpack with him because I hope that breakfast has an influence.

**EXPERIMENT CONSTRUCTION**

As extensively reported (Tschirgi, 1980; Schauble, 1996), the strategy yielding optimal performance (as it requires the least possible number of experiments and cognitive energy to discover regularities) in this type of task is the “vary one thing at a time” strategy (VOTAT), also known as the “control of variables strategy” (CVS, Chen & Klahr, 1999). Learners applying this strategy construct experiments in which they vary the variable under study while keeping all other variables constant. A number
of learners actually referred to this strategy during the think-aloud sessions, see the excerpts below.

<table>
<thead>
<tr>
<th>Learner 39</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Experiment 6 and 7:</td>
</tr>
<tr>
<td>Protocol:</td>
</tr>
<tr>
<td>Nou waarom ik trouwens elke keer hetzelfde rijtje doe is omdat ik ehm, omdat je dan gewoon kan zien welke... Omdat er eentje anders is en als de tijd dan anders is dan ligt het dus aan die ene.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Learner 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Experiment 2 and 3:</td>
</tr>
<tr>
<td>Protocol:</td>
</tr>
<tr>
<td>OK nou, dan ga ik nu kijken wat er gebeurt als hij op zijn mountainbike naar school zou gaan. En in hoeverre hij dan dus trager is. Verder heb ik alle keuzes hetzelfde dus ik blijf bij 't brood thuis opeten, weinig boeken, sportschoenen en rugzak want zo kan ik dan ontdekken in hoeverre het uitmaakt met watvoor fiets hij gaat.</td>
</tr>
</tbody>
</table>

Learners not using a VOTAT-like strategy conduct experiments to test a variable without taking care of keeping the other variables constant, see the excerpt below.
Models of Inductive Learning

<table>
<thead>
<tr>
<th>Learner 21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiments 7:</td>
</tr>
<tr>
<td><img src="apple.png" alt="apple" /> <img src="kiwi.png" alt="kiwi" /> <img src="wine.png" alt="wine" /> <img src="bread.png" alt="bread" /> <img src="pear.png" alt="pear" /></td>
</tr>
<tr>
<td>Pred.: 6.45</td>
</tr>
</tbody>
</table>

**Protocol:**

[...] Ga ik toch maar weer even voor peer, karnemelk, 7up, wit brood en een ui omdat ik dit rijtje bijna identiek heb aan rijtje 7 en ik wil weten of het uitmaakt of er een verschil is in prijsklasse met de peer en de appel. [...] 6.65 had ik toen en dan krijg ik nu weer 6.45 schat ik. Then I'll go for pear, buttermilk, 7-up, white bread and an onion because I have this row almost equal to row 7 and I want to know if it matters if there is a difference in price between apple and pear. It was 6.65, and then I'll get another 6.45 now, I guess.

<table>
<thead>
<tr>
<th>Learner 21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiments 9:</td>
</tr>
<tr>
<td><img src="apple.png" alt="apple" /> <img src="kiwi.png" alt="kiwi" /> <img src="wine.png" alt="wine" /> <img src="bread.png" alt="bread" /> <img src="pear.png" alt="pear" /></td>
</tr>
<tr>
<td>Pred.: 6.45</td>
</tr>
</tbody>
</table>

**Protocol:**

En dat is goed. [...] And that is correct. [...] The learner performs experiment 8 but makes an error constructing it, discovers the error immediately, and constructs experiment 9 according to the reported think-aloud excerpt.

Based on these experiments, one might infer that the learner knows that conducting a VOTAT experiment leads to interpretable data. However, directly after this experiment, the learner conducts an experiment that, according to the think-aloud protocol, is not directly related to another experiment and does not yield new knowledge.

<table>
<thead>
<tr>
<th>Learner 21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiments 10:</td>
</tr>
<tr>
<td><img src="apple.png" alt="apple" /> <img src="kiwi.png" alt="kiwi" /> <img src="wine.png" alt="wine" /> <img src="bread.png" alt="bread" /> <img src="pen.png" alt="pen" /></td>
</tr>
<tr>
<td>Pred.: 6.55</td>
</tr>
</tbody>
</table>

After the construction of this experiment, the learner seems a bit lost, and after some reasoning about the conducted experiments, a new experiments is constructed. During the construction of this experiment, the learner states:
Just before Experiment 11:

<table>
<thead>
<tr>
<th>Protocol:</th>
<th>Outcome:</th>
<th>Pred.:</th>
</tr>
</thead>
<tbody>
<tr>
<td>[...] Ja, dan twee artikelen anders. [...] Yes, then two products different. [...]</td>
<td>6.55</td>
<td>6.65</td>
</tr>
</tbody>
</table>

Experiment 11:

<table>
<thead>
<tr>
<th>Protocol:</th>
<th>Outcome:</th>
<th>Pred.:</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.55. Hm.. ehm, ja... [...] volgens mij maakt het niks uit, de drank en de prei.</td>
<td>6.55. Hmmm, cuhm, yes... [...] I don’t think that it matters, the beverage and the leek.</td>
<td></td>
</tr>
</tbody>
</table>

So, even after conducting a correct VOTAT experiment-pair (Experiment 7 and 9), the learner constructs two invalid VOTAT experiments. Even more, the learner infers incorrectly that “the beverage and the leek” do not influence the score.

Gathering data to test the effect of a variable specified in the current hypothesis is the main purpose of the experiment construction part. As this process has access to all previously constructed experiments, it enables the experiment construction process to check the history of constructed experiments to prevent duplication. For example, assume that the current hypothesis is about the effect of the different types of shoes. Then, the experiment construction process has either to create or to find experiments that can be used for comparing the effect of the shoes. As the levels of the other variables are not constrained, one of the “old” experiments can be used as initial experiment requiring the construction of only one new experiment if the other experiment is used as contrast. With respect to prior knowledge, as the Peter-task guides the learner in the experiments that need to be conducted (e.g., by presenting the learner with predefined discrete levels of the variables), prior knowledge does not play a role in the construction of experiments. That is, in tasks where the learner has more liberty in constructing experiments, e.g., by determining which variables are tested, prior knowledge can influence this selection and also determine which levels of these variables are actually tested and compared.

Because of the relatively simple form of the hypotheses in the Peter-task, the main purpose of the hypothesis is to guide the experiment construction process by iden-
Models of Inductive Learning

tifying which variable is the one under study. Therefore, the process in the models can be summarized as: $M_2(E + H) \rightarrow E_i$.

Knowledge Derivation

After two experiments have been constructed according to the above described procedures, (or after the learner “recycled” one experiment and constructed one new experiment) the experiments are compared. This comparison tests the correctness of the current hypothesis by deriving the effect of the variable mentioned in the hypothesis. Although the conclusions from other experiments might play a role in the derivation of knowledge in more complex task settings (e.g., where the type of relations discovered in earlier experiments influences the derivational process), this does not seem to interfere in the Peter-task. However, some learners showed confirmation bias effects (Klayman & Ha, 1987) caused by prior knowledge. If their beliefs about the effect of a variable are too strong, the results of the experiments are sometimes ignored in favor of the original assumption. Although incorporating confirmation bias in the models presented later would be a straightforward exercise, the confirmation biases related effects in the data are too divers and idiosyncratic to warrant a precise implementation. Therefore, the current implementation does not utilize $PK$ as data, yielding $M_3(H + E) \rightarrow C_i$ as the summary of this subtask.

Nevertheless, prior knowledge plays an indirect role in the knowledge derivation process. The hypothesis that is tested is based on an evaluation of prior knowledge. If there is a discrepancy between the experimental outcomes and the hypothesis, this discrepancy might trigger the conclusion that the effect of this variable needs to be tested further. Although in the models presented in this chapter the noticed discrepancies are only tested at the end of processing, this could also occur directly after noticing the discrepancy. However, most learners seem to give priority to testing for all main effects above testing the discrepancies. Even though delaying the testing of the discrepancy increases the change of forgetting about the necessity to test the found discrepancy.

\footnote{Which can be contrasted to domains that require more elaborate processing of the current hypothesis in the construction of experiments, for example, if there is no direct match between hypothesis and experimental variables.}
DETERMINING WHEN TO STOP

In most inductive learning studies, the stop criterion is deliberately kept opaque. Although learners are instructed “to discover how X works”, or “to discover what happens if Y and Z”, they are not told what level of detail is necessary. Learners have to decide for themselves when to stop experimenting. This also holds for the Peter-task. Although learners quickly infer that all five variables have to be tested (prescribing a minimum number of experiments, \( E \)), after testing these five variables a learner has to decide if other effects have to be tested too.

This decision is influenced by the type of discoveries that were made during the testing of the five variables and the prior knowledge a learner has. If during the testing of the main effects of the five variables other potentially interesting effects surfaced as conclusions of the related experiments (\( C \), e.g., as discrepancies between prior knowledge, conducted experiments and conclusions), the learner might decide to pursue these further. With respect to the prior knowledge (\( PK \)), if a learner recalls a situation in which a choice of a level for one variable influenced the effect of another variable (e.g., that biking on a racing-bike requires both hands, and that therefore eating breakfast while biking might negate the effect of the racing-bike), hypotheses constructed on the basis of prior knowledge might be pursued delaying the stop decision.

But, most importantly, if a hypothesis (\( H \)) is still being tested or if there are still hypotheses waiting to be tested, deciding to stop is premature.

CHECKING PRIOR KNOWLEDGE

Although not present as a separate entry in the four subtasks of Table 4.2, the prior knowledge plays an important role in inductive learning. In the models presented later this chapter, the prior knowledge determines the hypotheses on the basis of which the learners start conducting their experiments. Moreover, if these experiments result in derivations incompatible with the prior knowledge, the model signals this incompatibility which might lead to further examination. Therefore, the type and presence of prior knowledge plays an important role in determining the resulting behavior.

We will present four models in the next section with two different bodies of prior
knowledge. One set in which prior knowledge is specified about the effect of bikes and breakfast, and another set without prior knowledge.

THE MODELS

In this section, a number of different models is presented that together show a range of behavior similar to human behavior. The general structure of all models is depicted in Figure 4.3. By enabling or disabling certain aspects of the model, different types of behavior emerge. The models are developed using ACT-R\(^5\) (Anderson, 1990; Anderson & Lebiere, 1998) version 5.0 (Lebiere, 2001). The main focus in these models are (the interactions between) the mechanisms that are underlying inductive learning behavior. Therefore, emphasis is on the symbolic level\(^6\).

The discussion and description of the four subtasks in the previous section was at a verbal level. Below, we discuss the subtasks at a more detailed level, directly related to the computational implementation of the subtasks.

\[ M_1(PK + C) \rightarrow H \]

Assuming that a learner knows that a hypothesis is most likely about the differences between two levels of a variable, the most straightforward method is to search for prior knowledge that represents previous encounters with those levels and to compare the related results. This process is implemented in the models presented in this chapter.

With respect to the influence of the conclusions based on the experiments, the model compares the outcome of the experiments with the hypothesis formulated earlier (see Knowledge derivation, discussed earlier). If the outcome and the hypothesis are inconsistent, the model notices a mismatch which lead to an examination of the differences between the prior knowledge and the current situation. This examination might result in the construction of a new hypothesis to "retest" the effect of the variable. However, less motivated learners might notice a discrepancy without actually testing its implications.

\[ M_2(E + H) \rightarrow E \]

Three implementations of \( M_2 \) will be presented. Two of these apply the VOTAT strategy by always selecting the first option of the variables

\(^5\)See Chapter 3, 21 for a short introduction to ACT-R.

\(^6\)Subsymbolic information is only utilized to have the model favor more recently constructed experiments over older experiments, see Chapter 2 for a model that relies more heavily on subsymbolic learning.
Figure 4.3. Outline of the inductive learning task models.
that are not under study. One of these does not "reuse" already conducted experiments, the other implementation does reuse old experiments. The third implements a non-VOTAT strategy, referred to as the "sloppy experimenting" mode, which has a tendency to select the first option but randomly chooses the other level in 25% of all cases.

\[ M_3(H + E) \rightarrow C_i \] The implemented \(M_3\) process examines the (last two) conducted experiments and tests the correctness of the specified hypothesis. If the knowledge derived from the experiments contradicts the hypothesis, the model notices this discrepancy by adding a reference to the current, falsified hypothesis and the experiments that led to the falsification. Later, the Hypothesis formulation subtask can use this knowledge to re-examine this unexpected effect which might lead to testing for interactions.

\[ M_4(PK + E + H + C) \rightarrow \text{stop} \] The implemented \(M_4\) tests if all experiments are conducted that are required by the task format, if all hypotheses are tested, and if all salient discrepancies between prior knowledge and the conclusion based on the experimental outcomes are solved. If these conditions are true, the model stops experimenting.

For some subtasks, multiple approaches can be derived for a single subtask, and the outcome of a subtask might be dependent on the availability of certain knowledge. Therefore, multiple combinations are possible. Table 4.3 presents these alternatives per subtask.

All these alternatives are modeler selected, that is, the modeler decides at the start of a model run which of the alternatives is available. With respect to the two alternatives for \(M_1\), the modeler can decide to provide the model with prior knowledge regarding some variables or none at all. The knowledge derivation subtask does not have any alternatives, as all learners in the Peter-task showed the ability to infer the effect of two levels of a variable from two experiments in which this variable was modified. However, some learners (incorrectly) induced an effect for a variable when the other levels were not kept constant. This is simulated by assuming that the knowledge derivation process examines two experiments constructed by other subprocesses solely for determining the effect of a specified variable. This way, this process is not influenced by variations in levels of other variables. As the experiment constructing processes
Table 4.3: The alternative mechanisms or outputs of the mechanisms per subtask

<table>
<thead>
<tr>
<th>Subtask</th>
<th>Alternatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis formation</td>
<td>$M_1$ Specification of variable <em>and</em> prediction  \</td>
</tr>
<tr>
<td></td>
<td>As above, but <em>without</em> a prediction  \</td>
</tr>
<tr>
<td>Experiment construction</td>
<td>$M_2$ VOTAT mode, reusing old experiments \</td>
</tr>
<tr>
<td></td>
<td>VOTAT mode, not reusing old experiments \</td>
</tr>
<tr>
<td></td>
<td>Sloppy experimentation mode  \</td>
</tr>
<tr>
<td>Knowledge derivation</td>
<td>$M_3$ No alternatives in the current task context \</td>
</tr>
<tr>
<td>Determining when to stop</td>
<td>$M_4$ Testing found irregularities \</td>
</tr>
<tr>
<td></td>
<td>Stop when five variables are tested  \</td>
</tr>
</tbody>
</table>

*Note*: the first two alternatives, although specified as “hypothesis formation” alternatives, are actually better thought of as two alternatives with respect to prior knowledge. That is, a prediction can only be constructed if prior knowledge is available.

have to take care of constructing appropriate (i.e., VOTAT) experiments, the incorrect inducing of knowledge is caused by the non-VOTAT experiment construction subtask ($M_2$).

When the ACT-R model is run, it mimics a think-aloud protocol by outputting “think-aloud like” information (c.f., model-tracing, see for example Anderson, 1990; Anderson, Boyle, Corbett, & Lewis, 1990; Jansweijer, 1988; Jansweijer, Elshout, & Wielinga, 1989) for the important subtasks in the inductive learning process. A complete run of the model is presented below. In this model, the following alternatives were selected (a) prior knowledge that leads to hypotheses (represented in the model’s think-aloud protocol by their numbers, e.g., “cf. 4 vs 1” in line 2 and discussed later), (b) VOTAT mode with reusing experiments, and (c) testing the irregularities that were found.

**Model 1**

1. Let’s see, a hypothesis for variable Bike...
2. Often (cf. 4 vs 1), a BikeRacing is faster than/to a BikeNormal, so that will be my hypothesis
3. Going to construct an experiment to test BikeRacing
4. Cannot find an experiment testing BikeRacing, so constructing a new experiment.
5. Experiment 1: $\text{add}$ $\text{add}$ $\text{add}$ $\text{add}$ $\text{add}$ 40

*In contrast to the actual Peter-task, which has 3 levels for the variable “bag”, the current model only has 2 levels. Inclusion of the third level would complicate the computational models and their discussion whereas it would not increase their explanatory power.*
Models of Inductive Learning

Experiment 2:  
I knew that it was faster
Let's see, a hypothesis for variable Breakfast...
Often (cf. 3 vs 4), a BreakfastAtHome is slower than/to a BreakfastOnBike, so that will be my hypothesis
Going to construct an experiment to test BreakfastAtHome
It seems that I can use experiment 1
Experiment 3:  
Hmmm, I thought it was Slower, but the experiments 1 and 3 indicate BreakfastAtHome is faster than/to Breakfastbik
Well, I've seen that effect before, but it is strange...
Let's see, a hypothesis for variable Bag...
Hmmm, no idea what a BagSchoolbag does
Hmmm, no idea what a BagBackpack does
Hmmm, I don't know anything about BagBackpack, so, no hypothesis.
Going to construct an experiment to test BagSchoolbag
It seems that I can use experiment 3
Experiment 4:  
I didn't know what the relation would be, the experiments 3 and 4 indicate Slower
Let's see, a hypothesis for variable Books...
Hmmm, no idea what a BooksAll does
Hmmm, no idea what a BooksSome does
Hmmm, I don't know anything about BooksSome, so, no hypothesis.
Going to construct an experiment to test BooksAll
It seems that I can use experiment 4
Experiment 5:  
I didn't know what the relation would be, the experiments 4 and 5 indicate Equal
Let's see, a hypothesis for variable Shoes...
Hmmm, no idea what a ShoesNormal does
Hmmm, no idea what a ShoesSport does
Hmmm, I don't know anything about ShoesSport, so, no hypothesis.
Going to construct an experiment to test ShoesNormal
It seems that I can use experiment 5
Experiment 6:  
I didn't know what the relation would be, the experiments 5 and 6 indicate Equal
Ok, I'm ready testing the five variables...
I've discovered that:
BikeRacing is Faster than/to BikeNormal
BreakFastAtHome is Faster than/to BreakfastOnBike
BagSchoolbag is Slower than/to BagBackpack

6  Experiment 2:  
7  I knew that it was faster
8  Let's see, a hypothesis for variable Breakfast...
9  Often (cf. 3 vs 4), a BreakfastAtHome is slower than/to a BreakfastOnBike, so that will be my hypothesis
10  Going to construct an experiment to test BreakfastAtHome
11  It seems that I can use experiment 1
12  Experiment 3:  
13  Hmmm, I thought it was Slower, but the experiments 1 and 3 indicate BreakfastAtHome is faster than/to Breakfastbik
14  Well, I've seen that effect before, but it is strange...
15  Let's see, a hypothesis for variable Bag...
16  Hmmm, no idea what a BagSchoolbag does
17  Hmmm, no idea what a BagBackpack does
18  Hmmm, I don't know anything about BagBackpack, so, no hypothesis.
19  Going to construct an experiment to test BagSchoolbag
20  It seems that I can use experiment 3
21  Experiment 4:  
22  I didn't know what the relation would be, the experiments 3 and 4 indicate Slower
23  Let's see, a hypothesis for variable Books...
24  Hmmm, no idea what a BooksAll does
25  Hmmm, no idea what a BooksSome does
26  Hmmm, I don't know anything about BooksSome, so, no hypothesis.
27  Going to construct an experiment to test BooksAll
28  It seems that I can use experiment 4
29  Experiment 5:  
30  I didn't know what the relation would be, the experiments 4 and 5 indicate Equal
31  Let's see, a hypothesis for variable Shoes...
32  Hmmm, no idea what a ShoesNormal does
33  Hmmm, no idea what a ShoesSport does
34  Hmmm, I don't know anything about ShoesSport, so, no hypothesis.
35  Going to construct an experiment to test ShoesNormal
36  It seems that I can use experiment 5
37  Experiment 6:  
38  I didn't know what the relation would be, the experiments 5 and 6 indicate Equal
39  Ok, I'm ready testing the five variables...
40
41  I've discovered that:
42  BikeRacing is Faster than/to BikeNormal
43  BreakFastAtHome is Faster than/to BreakfastOnBike
44  BagSchoolbag is Slower than/to BagBackpack
45 BooksAll is Equal than/to BooksSome
46 ShoesNormal is Equal than/to ShoesSport
47
48 Hmm, there was something strange... Let’s check that
49 The experiments 1 and 3 did not match the expectation that BreakFastAtHome is Faster than/to
Breakfastbike...
50 The memories on which I based the hypothesis are Mem4 and Mem3
51 Maybe the effect is different because of the type of Bike...
52 Let’s test it by constructing the same experiments but now with the other bike-level...
53 Experiment 7: BikeRacing BreakfastAtHome 50
54 Experiment 8: BikeNormal BreakfastAtHome 50
55 Yes, the effect of BREAKFAST depends on BIKE!
56 If BikeRacing: BreakFastAtHome is Faster than/to BreakfastOnBike
57 If BikeNormal: BreakFastAtHome is Equal than/to BreakfastOnBike
58
59 No (other) irregularities... Ready!

As can be seen in this output, the model iterates over all five variables, searches for
a hypothesis per variable, and if found, tests that hypothesis by constructing experi­
ments and comparing the outcome of those experiments with the found hypothesis.
The experiments constructed by the model and the conclusions derived from those
experiments are presented in Table 4.5.

Based on these experiments, the model derived all main-effects (starting at lines 1,
8, 15, 23, and 31) and because the discovered effect for the variable breakfast was
unexpected (line 13), it reexamined this effect after it tested the effect of the five
variables. This reexamination (starting at line 48) involves comparing the context of
the prior knowledge leading to the original falsified hypothesis and the experiments.
The knowledge available to the model is:

<table>
<thead>
<tr>
<th>Memory 3:</th>
<th>Memory 4:</th>
</tr>
</thead>
<tbody>
<tr>
<td>bike</td>
<td>BikeNormal</td>
</tr>
<tr>
<td>breakfast</td>
<td>breakfastbike</td>
</tr>
<tr>
<td>bag</td>
<td>nil</td>
</tr>
<tr>
<td>books</td>
<td>nil</td>
</tr>
<tr>
<td>shoes</td>
<td>nil</td>
</tr>
<tr>
<td>outcome</td>
<td>20</td>
</tr>
<tr>
<td>bike</td>
<td>BikeNormal</td>
</tr>
<tr>
<td>breakfast</td>
<td>BreakFastAtHome</td>
</tr>
<tr>
<td>bag</td>
<td>nil</td>
</tr>
<tr>
<td>books</td>
<td>nil</td>
</tr>
<tr>
<td>shoes</td>
<td>nil</td>
</tr>
<tr>
<td>outcome</td>
<td>30</td>
</tr>
</tbody>
</table>
Table 4.5: Model 1: Experiments and conclusions of a learner with prior knowledge, a VOTAT experiment construction strategy and testing of unexpected effects.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Bike</th>
<th>Breakfast</th>
<th>Bag</th>
<th>Books</th>
<th>Shoes</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Racing</td>
<td>At home</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>Normal</td>
<td>At home</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>Racing</td>
<td>Bike</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>55</td>
</tr>
<tr>
<td>4</td>
<td>Racing</td>
<td>Bike</td>
<td>Backpack</td>
<td>All</td>
<td>Normal</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>Racing</td>
<td>Bike</td>
<td>Backpack</td>
<td>Some</td>
<td>Normal</td>
<td>50</td>
</tr>
<tr>
<td>6</td>
<td>Racing</td>
<td>Bike</td>
<td>Backpack</td>
<td>Some</td>
<td>Sport</td>
<td>50</td>
</tr>
<tr>
<td>7</td>
<td>Normal</td>
<td>At home</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>50</td>
</tr>
<tr>
<td>8</td>
<td>Normal</td>
<td>Bike</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable(s)</th>
<th>Constraint</th>
<th>Level conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main effects:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Bike</td>
<td></td>
<td>“racing” is faster than “normal”</td>
</tr>
<tr>
<td>2 Breakfast</td>
<td></td>
<td>“at home” is faster than “bike”</td>
</tr>
<tr>
<td>3 Bag</td>
<td></td>
<td>“book bag” is slower than “backpack”</td>
</tr>
<tr>
<td>4 Books</td>
<td></td>
<td>“all” is equal to “some”</td>
</tr>
<tr>
<td>5 Shoes</td>
<td></td>
<td>“normal” is equal to “sport”</td>
</tr>
<tr>
<td>Interactions/Conditional effects:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6a Breakfast Bike = “racing”</td>
<td>“at home” is faster than “on bike”</td>
<td></td>
</tr>
<tr>
<td>6b Breakfast Bike = “normal”</td>
<td>“at home” is equal to “on bike”</td>
<td></td>
</tr>
</tbody>
</table>

Note that although the format of the information is similar to the format used in the Peter-domain, this prior knowledge represents information acquired outside the Peter-task. This prior knowledge reflects that the learner knows about a situation in which having breakfast on a normal bike is faster than having the breakfast at home, and then biking to school on a normal bike. Note that the actual values represented in this overview are not equal to outcomes that are associated with these levels in the Peter-task. This reflects that the prior knowledge has a different source than the present task. When the model tries to discover what causes the difference between the prior knowledge and the constructed experiments, it discovers that the type of bike differs (line 51). Consequently, the model constructs two new experiments comparable to the earlier constructed experiments for the breakfast-test. However, in these experiments the type of bike is changed (lines 52-54). By comparing the effect of the breakfast levels given the levels of the bike the model discovers that the effect of breakfast depends on the type of bike used (lines 55-57), in other words, it
discovering the interaction.

Note that this model has discovered the total set of underlying effects, even given that it did not cover the complete experiment space neither generate all possible outcomes.

**MODEL 2**

If the same model is run without access to prior knowledge about the variables, no unexpected results are discovered. Therefore, after the model has constructed the experiments necessary to test the five variables, it stops before it has discovering the total set of effects. Table 4.6 presents the conducted experiments and conclusions. Note that this model does not discover the interactions between bike and breakfast. Because the model does not have access to prior knowledge, the model does not (and cannot) note any inconsistencies and does therefore not discover or test for any interactions.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Bike</th>
<th>Breakfast</th>
<th>Bag</th>
<th>Books</th>
<th>Shoes</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Racing</td>
<td>At home</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>Normal</td>
<td>At home</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>Racing</td>
<td>Bike</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>55</td>
</tr>
<tr>
<td>4</td>
<td>Racing</td>
<td>Bike</td>
<td>Backpack</td>
<td>All</td>
<td>Normal</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>Racing</td>
<td>Bike</td>
<td>Backpack</td>
<td>Some</td>
<td>Normal</td>
<td>50</td>
</tr>
<tr>
<td>6</td>
<td>Racing</td>
<td>Bike</td>
<td>Backpack</td>
<td>Some</td>
<td>Sport</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable(s)</th>
<th>Constraint</th>
<th>Level conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main effects:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Bike</td>
<td>“racing” is faster than “normal”</td>
<td></td>
</tr>
<tr>
<td>2 Breakfast</td>
<td>“at home” is faster than “bike”</td>
<td></td>
</tr>
<tr>
<td>3 Bag</td>
<td>“book bag” is slower than “backpack”</td>
<td></td>
</tr>
<tr>
<td>4 Books</td>
<td>“all” is equal to “some”</td>
<td></td>
</tr>
<tr>
<td>5 Shoes</td>
<td>“normal” is equal to “sport”</td>
<td></td>
</tr>
</tbody>
</table>

**MODEL 3**

If Model 2 would have used the experiment construction method that does not reuse old experiments, the derived knowledge would be similar but the model would have
constructed four more experiments as can be seen in Table 4.7.

Table 4.7: Model 3: Experiments and conclusion of a learner without prior knowledge, but with a VOTAT experiment construction strategy without reusing already conducted experiments and with testing of unexpected effects.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Bike</th>
<th>Breakfast</th>
<th>Bag</th>
<th>Books</th>
<th>Shoes</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Racing</td>
<td>At Home</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>Normal</td>
<td>At Home</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>Racing</td>
<td>At Home</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>40</td>
</tr>
<tr>
<td>4</td>
<td>Racing</td>
<td>Bike</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>55</td>
</tr>
<tr>
<td>5</td>
<td>Racing</td>
<td>At Home</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>40</td>
</tr>
<tr>
<td>6</td>
<td>Racing</td>
<td>At Home</td>
<td>Backpack</td>
<td>All</td>
<td>Normal</td>
<td>35</td>
</tr>
<tr>
<td>7</td>
<td>Racing</td>
<td>At Home</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>Racing</td>
<td>At Home</td>
<td>Book bag</td>
<td>Some</td>
<td>Normal</td>
<td>40</td>
</tr>
<tr>
<td>9</td>
<td>Racing</td>
<td>At Home</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>40</td>
</tr>
<tr>
<td>10</td>
<td>Racing</td>
<td>At Home</td>
<td>Book bag</td>
<td>All</td>
<td>Sport</td>
<td>40</td>
</tr>
</tbody>
</table>

Variable(s) | Constraint | Level conclusions
------------|-------------|-------------------
Main effects:
1 Bike | “racing” is faster than “normal”
2 Breakfast | “at home” is faster than “bike”
3 Bag | “book bag” is slower than “backpack”
4 Books | “all” is equal to “some”
5 Shoes | “normal” is equal to “sport”

MODEL 4

In spite of the differences in behavior of the discussed models, they all contain only “correct” knowledge and strategies. Regardless of the limited scope of discovered information in the last two models, none of the subprocess can be blamed for not discovering all relations. However, some learners tested in the before-mentioned studies did behave erratically. Tables 4.8 and 4.9 (split into two because of larger number of conducted experiments) shows the experiments constructed by a model that violates the VOTAT principle. (Note that this model had access to the same prior-knowledge as Model 1, Table 4.5.)

As can be seen in Table 4.9, regardless of the incorrect experimenting strategy and therefore also the incorrect knowledge derivation strategy (as the model derives the effect of the variable under study from experiments that were constructed without
paying attention to the “other” variables), this model correctly derives knowledge about the interaction.

**Variations of these Models**

Obviously, these models reflect only a limited subset of behavior shown by learners in settings like the Peter-task. The main limitation of these models is the lack of heuristics in the determining when to stop subtask ($M_4$). In the currently presented models, the stop subtask is straightforward: as soon as the effect for all variables is tested, and no knowledge about possible other effects is available, the model stops. However, think-aloud protocols of learners in the Peter-task show different types of heuristics that influence the decision to stop. For example, learners actively engage in reasoning about the duration and complexity of examining all possible (first-order) interactions after discovering the first interaction. For example, the learner in the excerpt below decides to stop experimenting after a judgement of complexity of continuing.

Table 4.8: Model 4: Experiments of a learner with prior knowledge, but without a VOTAT experiment construction strategy, without reusing already conducted experiments and with testing of unexpected effects.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Bike</th>
<th>Breakfast</th>
<th>Bag</th>
<th>Books</th>
<th>Shoes</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Racing</td>
<td>At Home</td>
<td>Book bag</td>
<td>Some</td>
<td>Sport</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>Normal</td>
<td>At Home</td>
<td>Book bag</td>
<td>All</td>
<td>Sport</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>Racing</td>
<td>At Home</td>
<td>Book bag</td>
<td>Some</td>
<td>Sport</td>
<td>40</td>
</tr>
<tr>
<td>4</td>
<td>Racing</td>
<td>Bike</td>
<td>Book bag</td>
<td>Some</td>
<td>Sport</td>
<td>55</td>
</tr>
<tr>
<td>5</td>
<td>Racing</td>
<td>At Home</td>
<td>Book bag</td>
<td>Some</td>
<td>Normal</td>
<td>40</td>
</tr>
<tr>
<td>6</td>
<td>Racing</td>
<td>Bike</td>
<td>Backpack</td>
<td>Some</td>
<td>Sport</td>
<td>50</td>
</tr>
<tr>
<td>7</td>
<td>Racing</td>
<td>At Home</td>
<td>Book bag</td>
<td>All</td>
<td>Sport</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>Racing</td>
<td>At Home</td>
<td>Book bag</td>
<td>Some</td>
<td>Sport</td>
<td>40</td>
</tr>
<tr>
<td>9</td>
<td>Racing</td>
<td>Bike</td>
<td>Book bag</td>
<td>Some</td>
<td>Normal</td>
<td>55</td>
</tr>
<tr>
<td>10</td>
<td>Racing</td>
<td>Bike</td>
<td>Backpack</td>
<td>Some</td>
<td>Sport</td>
<td>50</td>
</tr>
<tr>
<td>11</td>
<td>Normal</td>
<td>Bike</td>
<td>Book bag</td>
<td>All</td>
<td>Normal</td>
<td>50</td>
</tr>
<tr>
<td>12</td>
<td>Normal</td>
<td>At Home</td>
<td>Book bag</td>
<td>Some</td>
<td>Normal</td>
<td>50</td>
</tr>
</tbody>
</table>
Models of Inductive Learning

Table 4.9: Model 4: Conclusions of a learner with prior knowledge, but without a VOTAT experiment construction strategy, without reusing already conducted experiments and with testing of unexpected effects.

<table>
<thead>
<tr>
<th>Variable(s)</th>
<th>Constraint</th>
<th>Level conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Bike</td>
<td></td>
<td>“racing” is faster than “normal”</td>
</tr>
<tr>
<td>2 Breakfast</td>
<td></td>
<td>“at home” is faster than “bike”</td>
</tr>
<tr>
<td>3 Bag</td>
<td></td>
<td>“book bag” is faster than “backpack”</td>
</tr>
<tr>
<td>4 Books</td>
<td></td>
<td>“all” is equal to “some”</td>
</tr>
<tr>
<td>5 Shoes</td>
<td></td>
<td>“normal” is slower to “sport”</td>
</tr>
</tbody>
</table>

Interactions/Conditional effects:

| 6a Breakfast Bike = “racing” | “at home” is faster than “on bike” |
| 6b Breakfast Bike = “normal” | “at home” is equal to “on bike” |

Note: Incorrect conclusions are printed in **bold**.

---

Learner 36

Just after experiment 10:

Protocol:
Ja er zit iets onderling...maar...nee, ik weet niet wat...dan zou ik nog 10 more combinations...
Yes, there is some relation...but...no, I don't know what...then I'll have to do like 10 more combinations...

This learner did construct the necessary experiments to derive the interaction effect – but was not able to compare the correct set of experiments to derive the correct conclusion. Directly after the above statement, the learner decides to stop experimenting; the perceived complexity seems too high to continue experimenting. Or, a different learner, after discovering the first interaction, and contemplating whether to test for other interactions:
Another example is that some learners use their knowledge about the possible outcomes of the task to slightly change their approach to the task and to start actively searching for the outcomes they had not encountered yet in the constructed experiments:

<table>
<thead>
<tr>
<th>Learner 5</th>
<th>Learner 20</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Protocol:</strong></td>
<td><strong>Protocol:</strong></td>
</tr>
<tr>
<td>Ik denk als ik dat zeker wil weten dat ik dan alle dinges uit moet rekenen en alle dingen apart bij mekaar moet doen en dan ben ik echt wel een uur bezig of zo. OK, wat ga ik nu doen, ik weet nou niet meer zo goed waar ik dan moet beginnen of ik wel opnieuw moet beginnen om het uit te zoeken of niet. Ik denk eigenlijk dat het wel goed is ja ik denk dat ik klaar ben. Ik ben klaar.</td>
<td>Ehm, [stilte] nou ik heb dus de 35 minuten gehad, ehm 40 nog niet, 45, 50 en 55. Ehm, dan wil ik nu nog weten wanneer hij er drie kwartier over fietst, nee 40 minuten. Euhm, [silence] I have had the 35 minutes, euhm, haven’t had 40, 45, 50 and 55. Euhm, so I want to now when it takes him three quarters of an hour, no, 40 minutes.</td>
</tr>
</tbody>
</table>

In the first of these two examples, the heuristic intervenes at a meta-cognitive “planning” level. The second example is a more traditional heuristic, in that it specifies a method to check whether the found results are similar to a known distribution. If not, then not all possible outcomes have been generated, and there is a chance that there are still undiscovered effects. For both types of heuristics, (prior) knowledge is required. In the first example, the learner has to have an idea about the amount of
time that the total task can reasonably take. In the second example, the learner has to remember the set of possible outcomes from the task instruction.

Given the large variation in heuristics that are used by learners, it is not feasible to implement all these. Moreover, the idiosyncratic nature of the heuristics hinder a generalizing conclusion of the associated effects. Nevertheless, the heuristics play an important role in determining when to stop and therefore in the characterization of the inductive learning process. This is a deviation from the inductive learning theory as sketched in the SDDS approach. The subtask associated with deciding when to stop is underspecified, especially when compared to the level of specification of the other subtasks.

**DISCUSSION**

The above described models illustrate that behavior associated with the inductive learning as triggered by the Peter task can be described using four subtasks derived from the SDDS theory. Although derived from the SDDS framework, there are some notable differences. In the SDDS framework, as is visible by the level of elaboration of the left side of Figure 4.1, the emphasis is on the hypothesis formation steps. In contrast with this emphasis, the hypothesis formation methods in the models presented above are relatively simple. Moreover, the output of these methods is determined by the prior knowledge available to the model. Given the relatively simple structure of the Peter-task, it is obvious that it is relatively simple to find the correct type of hypothesis (in contrary to the BigTrak-task). Although the interactional effects in this task can be represented in a relatively simple manner as well (i.e., as conditional effects), the Peter-task is considered a real and difficult “discovery” task by learners and the less than perfect empirical results (see the next chapter) show that the task was not trivial.

Based on a task-analysis and the implemented models, variation in behavior between learners is explained in terms of different methods of experiment construction, different prior knowledge, and stop-heuristics. This shows the importance of these factors in the inductive learning process. Moreover, it also illustrates that the emphasis of the SDDS theory on the construction of hypotheses limits the applicability of this framework to domains in which the main task is to discover the type or form of a hypothesis (e.g., BigTrak). In contrast to those type of tasks, learning in Peter-like
task focuses on experiment generation and the application of a stop criterion. As was argued in this chapter, both the generation of experiment and the decision to stop experimenting are influenced by the assumptions learners have about the task and domain.

Obviously, there is great variability in the assumptions the learners have about the task, yielding variability in behavior. This effect is not only observable during the construction of experiments, but also influences the decision about when to stop experimenting. As argued, learners appear to experiment until they have manipulated all variables at least once and have solved any inconsistencies between their assumptions about the effect of the variables and the observed outcomes.

This is where the difference between an optimal rational solution and a more graded rational solution as observed in the learners' behavior is most obvious. The most rational solution to tasks like the Peter-task (which has 48 unique experiments) is to conduct all possible unique experiments. Then, by means of deduction, the learner can derive all relations that hold in the task under study. Learners in the Peter-task have not been given a clear “when-to-stop-experimenting” instruction. Therefore they have to judge for themselves when sufficient experiments have been conducted. A balance has to be found for this decision between discovering all possible relations and consuming too much resources. In other words, on the one hand, a learner tries to discover as much as possible, on the other hand, the learner also tries to limit the usage of resources. This behavior is often described as bounded rationality (Simon, 1957). (See Simon, 1991 for a comparison between “bounded” rationality and “pure” rationality). Because of this, effects of prior knowledge and subjective interpretation of the task play an important role in what behavior is shown by learners in this inductive learning task.

---

8Actually, there is no completely rational approach to this or similar tasks, as it always remains a possibility that the next to be conducted experiments gives new, not yet incorporated information. For example, the simulation might contain a training-effect along the lines of Peter biking faster if he biked to school at least 100 times.
CHAPTER 4.

MEASURES OF INDUCTIVE LEARNING QUALITY

Abstract

In this chapter we argue that two often used measures of the quality of inductive learning, comprehension score and proportion of experiment space covered, are subject to spurious influences of prior knowledge and subjective interpretation of the task. A new measure is proposed that is based on consistency between the observations made during learning and the post-test scores. It is shown that this measure is more stable over tasks than comprehension score and proportion experiment-space covered.

There are several approaches to measuring the quality of learners performing inductive learning tasks. The simplest approach is to evaluate the knowledge after performing inductive learning. How much is learned (e.g., percentage of relations that were correctly discovered) is a measure of the quality of the discovery behavior. However, as we discussed in the previous chapter, this measure is indirect and subject to the influence of other factors than the quality of inductive learning. For example, prior knowledge may produce a high “discovery” score even if it is combined with poor discovery behavior. Or, individual differences in the interpretation of the goal of the inductive learning task may result in differences in performance. In both cases measuring acquired knowledge does not reflect the quality of inductive discovery learning.

Parts of this and the previous chapter are based on Wilhelm et al. (in press) and Hulshof et al. (in press).
Another approach to measure inductive learning is to focus on the process of inductive learning, and to measure learning quality on the the subprocesses that underly inductive learning. In the previous chapter, we identified and discussed four sub-tasks of inductive learning. In the current chapter, we discuss existing measures of inductive learning in terms of these subtasks. After discussing these measures, we propose a new measure which is based on an assessment of the quality of one of the subprocesses.

**Measuring the Quality of Performance**

The difficulty is to measure the contribution and quality of the different sub-tasks in the discovery process and more specifically, which part is responsible for (sub)optimal behavior. Often, the approach taken is to rely on think aloud procedures (Chi, 1997; Ericsson & Simon, 1993; Van Someren, Barnard, & Sandberg, 1994) to identify what drives performance. However, another approach is to base analyses on, for example, reaction time data or human-task interaction data (Hulshof, 2001). We will refer to the first as qualitative data, and to the latter as quantitative data. In the next two paragraphs, we will discuss qualitative and quantitative measures that are traditionally used in the context of discovery learning.

**Measures Based on Qualitative Data**

When the think aloud method is used to gather data, a learner is asked to verbalize all thoughts during problem solving. By means of analyzing think aloud data, more insight in the relative contribution of the different phases can be gained. However, in the context of the inductive learning tasks presented in this thesis, think aloud data without prompting is often of a relatively low quality, especially with respect to the stating of hypotheses and intermediate results derived from experiments (see for a discussion on this topic Kuhn et al., 1995, p155, or Thinking Aloud, 2001). The ideal situation from a researchers point of view would be one in which the learners comment on each experiment. This led researchers to use prompting questions (e.g., Schauble, 1990; Niewold, 1998; Wilhelm, 2001). By means of these prompting questions, learners are asked explicitly to state their current hypotheses and expectations at fixed intervals or events. Niewold (1998) reported that asking directed questions (e.g., “What are you trying to discover?”) did not interfere with the score on measures
used to assess the quality of the inductive learning process. However, it is not obvious that asking questions does not change the discovery behavior at all, as Niewold also reported effects of prompting on other behavioral measures. Moreover, in other research on the use of prompting, an increase in performance was found (Berry & Broadbent, 1990).

Another issue associated with think aloud protocols is that the data collection and analysis of think aloud protocols is associated with enormous effort (cf. Hulshof, 2001), especially in situations where numerous protocols are necessary to derive generalities (cf. Prins, 2002). Therefore, verbal data like think-aloud protocols are often supported by (additional) behavioral data about learner actions. Since these actions can often be recorded automatically, it is easy to apply numerical measures to these data.

**Measures Based on Quantitative Data**

Besides qualitative descriptions of discovery behavior based on think aloud protocols, research often includes quantitative data reflecting the outcomes of behavior during the discovery process.

An often used quantitative data based measure that is supposed to reflect the outcomes of the discovery process is the “comprehension score” (Schauble, 1990). This score is based on the learner’s performance on sets of comprehension questions, asked in between experimental sessions or during a post-test. Whereas Schauble, Klopfer, and Raghavan (1991) calculated the comprehension score simply based on the (unweighted) number of correctly stated relations, Wilhelm (2001) and Niewold (1998) use a more elaborate, weighted scheme (see Table 5.1, page 86). Yet another classification system was used by Hulshof (2001). He classified the learners in the Peter-task into one of four different nominal levels, based on the number of correctly answered post-test questions.

Another measure, used to quantify the behavior of learners in a discovery task, is based on the number of (distinct) experiments conducted by learners (e.g., Klahr & Dunbar, 1988; Kuhn, Schauble, & Garcia-Mila, 1992; Schauble, 1990, 1996; Wilhelm, 2001). As this measure is used to describe how well the learners covered the total experiment space, the number of (distinct) experiments is divided by the total number of distinct experiments defined by the task. This experiment-space com-
pleteness score is presented as a measure of the thoroughness of scientific discovery learning.

**ON THE VALIDITY OF EXISTING MEASURES**

However, the validity of both the comprehension score and the proportion of experiment-space covered can be questioned with regard to the extent to which they measure the quality of discovery behavior.

**THE COMPREHENSION SCORE**

With regard to the comprehension score measure, the underlying idea is that a perfect learner should discover all results. However, a methodological problem with measures like this is that the number and type of relations in the domain is often not easily quantifiable. For example, should an interaction count as two main effects ($A_1 > A_2$ if $C = 1$ and $A_1 < A_2$ if $C = 2$), making it possible to weight the score for a partly correct answer, or should the interaction be seen as an inseparable construct yielding an all-or-none score? Or, similarly, should a learner be credited for correctly discovering that there is no effect of changing a particular variable to another level?

Another, more fundamental problem is associated with the assumption that perfect inductive discovery learning leads to the discovery of all possible effects. As argued in the previous chapter, it is often not reasonable to expect a learner to conduct all possible experiments. Therefore, a learner might make a valid decision to stop experimenting before all existing relations have been discovered. However, if the comprehension score is used as a measure of the quality of inductive learning, it automatically follows that discovering more effects is always better, ignoring the possibly valid reasons of a learner to stop experimenting. It is, for example, possible that a learner does not conduct experiments in the experiment space region where some relations can be discovered because the learner's prior knowledge suggests that the important relations are to be found in different regions. This distinction is also present in the work of Van Joolingen and De Jong (1997). They distinguish between two subspaces within the complete experiment-space. If derivation is based on experiments sampled from the “target conceptual model” space, this process would lead to a correct set of hypotheses. The actual subspace that learners consider during discovery learning is called the “effective learner search space”. Only if the “effective
Measures of Inductive Learning Quality

learner search space" overlaps the target conceptual model space, a learner is able to derive all hypothesis.

This is also illustrated in a comparison between the models as presented in the previous chapter (Model 1 vs Model 2 (or 3) vs Model 4). As Model 1 discovers all possible effects, this model gets the maximum comprehension score. Models 2 and 3 will both score lower, as the behavior of these models did not lead to the discovery of the interaction effects. However, given the commonly used scoring schemes for the comprehension score, Model 4, based on erroneous inductive learning processes, will score higher than Model 2 and 3 as this model did discover the interaction effects. (This is because the interaction effects contribute stronger to the comprehension score than the one wrong main effect.) However, the experiment construction mechanism of Model 4 is clearly imperfect. That the model discovers the correct effect by chance is no problem for a simple performance score, but a measure of the quality of the inductive learning process should not be biased by this type of accidental discoveries.

There is also a more theoretical reason why the comprehension score fails as a good measure of the quality of inductive learning. If one contrasts inductive learning to deductive learning, the latter requires that one searches for all possible experiments that could lead to deduction as a deductive learner should not make inferences about non-observed states. However, the nature of inductive learning implies that a learner induces hypotheses that hold over a broader range of experiment states than the observed states. Therefore, deciding not to search parts of the experiment space might be a decision that is completely valid in an inductive learning setting, even if that leads to impoverished knowledge. Therefore, although the comprehension score described above does measure certain aspects of discovery behavior, its validity is questionable.

THE PROPORTION OF EXPERIMENT-SPACE COVERED

The quantitative measure reflecting the proportion of experiment-space covered is often used to illustrate the (lack of) thoroughness of learners' discovery behavior. However, the implied correlation between the proportion score and the quality of the discovery process does not hold when it is not by definition considered suboptimal if a learner stops before all experiments have been conducted.

A measure used by Glaser, Schauble, Raghavan, and Zeitz (1992) illustrates this prob-
Modeling Inductive Learning

In this study, differences in “percentage of minimal required evidence” is reported for three types of tasks. This measure reflects the extent to which the learner conducted the necessary experiments to derive all possible relations. However, learners do not know what part of the experiment space contains the most informative relations. So, without more information about the discovery behavior of a learner, a negative interpretation of this measure would be that it just measures the amount of luck a learner had in selecting the right experiment space regions.

Another issue related to the proportion of experiment-space searched are duplicate experiments. A duplicate experiment is often considered unnecessary, and is interpreted negatively. However, two remarks can be made with regard to this negative interpretation. First, the learner does often not know whether the domain contains a stochastic or dependent (e.g., training effects) component. An appropriate test for stochasticity or dependency is repeating an experiment leading to a duplication. Second, an experiment that has been conducted a while ago might not be readily available to the learner. For example, because the declarative knowledge representing the experiment has decayed too much causing retrieval to be impossible. In both situations, constructing experiments which have already been conducted might be an appropriate action.

Based on the above reasoning, statements as presented in (Kuhn et al., 1992) are overstating the irrationality or suboptimality of the learners’ behavior. Kuhn et al. (1992, p.303, emphasis added) state: “Subjects generated only a fraction of the potential evidence, and this they did inefficiently. Of the 48 possible unique boats [experiments], subjects build [constructed] an average of 18.9 (39.4%, with a range across subjects from 27% to 56%).”. Here, “inefficiency” refers to duplicate experiments constructed by the learners.

The issue of duplicate experiments also plays a role when Model 2 and Model 3 presented in the previous chapter are compared. The only difference between Model 2 and Model 3 is that Model 3 does not reuse old experiments, but constructs a new “base experiment” for every variable it tests. Although this does take more effort in terms of mouse-clicks, this behavior is not incorrect. Although these duplicate experiments might seem redundant, implying mediocre inductive learning behavior is premature as long one does not know why these duplicate experiments have been constructed.
PRINCIPLED MEASURES

As discussed above, both the comprehension score and the proportion of experiment-space covered are not optimal to assess the quality of the inductive learning process. Therefore, an alternative measure is necessary that is not influenced by the issues raised previously. Given the resource-expensiveness of analyzing think aloud protocols, a measure based on log-file data is preferable over a measure based on qualitative data.

In the previous chapter, the four subtasks (constructing hypotheses, constructing experiments, evaluating hypotheses, determining when to stop) as defined by the SDDS theory have been implemented to construct computational models of inductive learning behavior. Based on an analysis of these subtasks, differences between good and bad discovery behavior can be accounted for. However, to assess this in empirical behavior, we have to have access to the products of the different subtasks. As argued before, there is no easy solution to access the hypotheses a learner might use during the discovery process. And even if we have access to the hypotheses a learner uses or verbalizes during the discovery process, it is difficult to assess whether these hypotheses are correctly derived from the learner’s prior knowledge ($M_1$). That is, it is (as in other fields of psychology) almost impossible to assess the prior knowledge ($PK$) a learner has about a domain without priming this prior knowledge and thereby influencing the inductive learning process itself.

As the hypothesis is the prime determinant of the experiment generation process in SDDS, assessing the quality of the generated experiments ($M_2$) is difficult. Because the prior knowledge is also the main determinant for the stop-criterion and the stop-criterion is not explicitly stated in the task instruction, which makes it subject to interpretation, the quality of this subtask ($M_4$) is not easily assessed.

This leaves the knowledge derivation ($M_3$) as main focus for measuring the inductive learning process. During this subtask, the learner derives new knowledge on the basis of conducted experiments. As argued above, prior knowledge and therefore also the current hypothesis might influence the derivational process as it might cause rash derivations or confirmation bias. However, both the product of this subtask, the conclusions stated during the post-test, and the input for this subtask, the experiments, are available for inspection. Therefore, we can easily reconstruct the effects of invalid
derivations and confirmation bias as these are reflected in invalid conclusions if one considers the experiments as infallible sources of evidence. For example, assume that we know that a learner constructed experiments in which all but one variable were kept constant. Changing the variable is reflected in the outcomes. However, if the learner does not mention this relation, we can assume that either the effect was not noticed, or that it might have been noticed but was not interpreted, probably because of confirmation bias. Either way, we can conclude that the inductive learning process of that learner is suboptimal: not enough attention was paid to an effect that was observable in the data.

Therefore, if aiming for an optimal quantifiable measure of the correctness of discovery behavior, the solution lies in measuring the inductive learning behavior while trying to keep the hard to measure subjective effects of prior knowledge and subjective interpretations under control. Method $M_3$ is most promising in this respect as all the input necessary for this method is easily available and the outcomes of this process are reported during the post-test (and in the think-aloud protocols if those have been recorded). Although this does not reveal what methods and prior knowledge are used by the learner, it does reveal if the learner is able to engage in correct discovery behavior given his or her capacities without being influenced by subjective properties like prior knowledge or interpretation.

However, if one does not specify any discovery strategies one cannot assess whether the behavior of the learner is consistent. Therefore, a minimal set of experimentation strategies has to be defined which are assumed to be available for the learners. As learners with similar background as the learners in the current study (Wilhelm, 2001; Niewold, 1998) showed to be consciously aware of the importance of the VOTAT or CVS method, we assume that the learners in the current study should be able to derive the correct knowledge from two correct experiments. For example, if a learner has two experiments available in which only one variable is changed, the underlying assumption of the consistency score requires that the user reports the effect that can be derived from those experiments during the post-test. With respect to interactions, if a learner saw earlier that changing a variable led to another effect than changing that variable later on, the VOTAT strategies should signal an inconsistency. Therefore, the consistency measure assesses the consistency of the overt behavior with the answers to the post-test scores assuming that the learners know that one has to test
for the effect of a variable by only manipulating that particular variable.

To test the viability of the consistency of behavior measure, we will apply this method to data from the Peter-task. The performance of the consistency measure will be contrasted with the performance of the original comprehension scores and the percentage experiment-space covered. If this measure is indeed less influenced by subjective properties, the score on this measure should be more stable over experiments than the previously used scores. Therefore, we will compare the stability of this consistency measure over two tasks with the stability of the comprehension score.

**PETER-TASK**

To compare measures for inductive learning tasks, we will again refer to the Peter-task as described in the previous chapter. First, we will present the scores on these tasks using the measures as reported in Wilhelm et al. (in press). Second, we will discuss these findings in the light of the issues raised earlier. Third, the consistency score will be operationalized and compared to the earlier measures.

**RESULTS**

We tested 15 subjects, all first-year psychology students at the University of Amsterdam. They participated in this study for course credits. The subjects were told that they could take as long as they wanted, and were given the opportunity to take notes. Post-test questions were used to assess what was learned. Per variable, the learner was asked “What did you discover concerning ...”. Only when the learner’s answer was ambiguous, clarification questions were asked.

As in previous studies (e.g., Wilhelm et al., in press; Schauble, 1990; Kuhn et al., 1992), only a fraction of the experiment space was searched. The average number of experiments conducted was 11 (SD = 5.78), the average number of unique experiments was 9.7 (SD = 3.53). The average proportion of the experiment space searched was 0.20 (i.e., 9.7 / 48, see Table 5.2).

With respect to the comprehension score, we used the same scoring scheme as used by Wilhelm et al. (in press). This scoring scheme weights the interactions by assigning more points to the interactions if completely reported and also allows for partial scoring of interactions. This leads to a maximum score of 18 points awarded if all
main effects and interactions are reported correctly. The actual scoring scheme used is depicted in Table 5.1. Based on the this scheme, the average comprehension score is 11.3 (SD = 2.4, ranging from 4 to 15, see Table 5.2).

Table 5.1: Points awarded for responses in the post-test in Wilhelm’s (2001, Table 2, p.37) study.

<table>
<thead>
<tr>
<th>Number</th>
<th>Statement</th>
<th>Correct</th>
<th>Without restricting</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$a_1 &gt; a_2 \ (if \ b_1)$</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>$a_1 &lt; a_2 \ (if \ b_2)$</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>$b_1 = b_2 \ (if \ a_1)$</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>$b_1 &gt; b_2 \ (if \ a_2)$</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>$c_1 = c_2$</td>
<td>2</td>
<td>n/a</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>$d_1 = d_2$</td>
<td>2</td>
<td>n/a</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>$e_1 &lt; e_2$</td>
<td>2</td>
<td>n/a</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>$e_1 &lt; e_2$</td>
<td>2</td>
<td>n/a</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>$e_2 = e_3$</td>
<td>2</td>
<td>n/a</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: Letters $a$ to $e$ refer to the five independent variables. The subscripted digits refer to the level of the variable selected. Variables $a$ and $b$ interact, $e$ has an effect for only one level and $c$ and $d$ are irrelevant. The maximum score is 18.

**DISCUSSION OF THE RESULTS**

With respect to the proportion of experiment space searched, the learners in the current study show a relatively low score, having generated only 20% of the possible evidence. The proportion of the experiment space searched depends on the number of experiments conducted. As no stop-rule or stop-time is specified at the start of the study, the learners have to decide for themselves whether or not to continue working after each experiment. As the experimental setting implies that there will be at least some effects, all learners at least try to discover the simple main effects. Hereto, most of the learners conduct the correct “vary one thing at a time” (VOTAT, Tschirgi, 1980; Schauble, 1996), also known as the “control of variables strategy (CVS, Chen & Klahr, 1999). (See Niewold, 1998 for similar reports on strategy use of Dutch first-year students.) Based on these strategies, learners are theoretically able to detect all main effects in 7 experiments (see Appendix 4.A). Knowing this, the actual number of unique experiments conducted by the learners (being 9.73), seems to reflect that

---

1 Using the unweighted scoring scheme as applied by (Kuhn et al., 1995) leads to qualitatively comparable results.
“just discovering the main effects” might come close to what learners actually tried to achieve.

This is also reflected in the comprehension scores. The maximum score using the scheme in Table 5.1 is 18. However, as some of these scores are related the practical maximum is in the order of 14. This is due to the “double scoring” for the interactions; point are given both for stating “the effect of $a$ depends on $b$” and for stating “the effect of $b$ depends on $a$”. Although this type of scoring is technically correct, it does not coincide with how the learners report the discovered effects. Learners often mention either the relations 1 and 2, or the relations 3 and 4 (Table 5.1), whereas in their think aloud protocol they mention the discovery of both sets of effects. Therefore, most learners score 14 points (2 x 2 points for the interactions, and 5 x 2 points for the main effects). Moreover, if a learner does not discover any of the interaction-related effects, the score is only 12 (i.e., one point for relation number 1 or 2 and 1 point for 3 or 4, and in total 10 points for the main effects and non-effects (relations 5 to 9)). As the actual average score is 11.3, this seems to indicate that learners chiefly discovered these main effects. Indeed, if we allow the learners to make one mistake in discovering the main effects, 12 of the 15 learners discovered solely these main effects (2 learners discovered parts of the interaction, 1 discovered only 3 of the possible main-effects).

As these measures are introduced as descriptors of the quality of the inductive learning process, the question is whether they reflect this accurately. As noted earlier, the inductive learning process is inherently uncertain. Learners do never know if the knowledge they have at a certain point in time is all knowledge that can be derived from the domain. If a learner in this domain did not think about interactions, but explored the experiment-space correctly given his prior knowledge and related assumptions, the quality of his behavior should be assessed given these ideas.

**Consistency Measure**

Having discussed the inappropriateness of the existing measures to assess the quality of the inductive learning behaviors we propose a new measure to assess the quality of the behavior in a more valid manner. We again turn to the post-test scores. However, instead of contrasting these with the maximum score, we compare them to the learner’s behavior. The rationale is that even if the learner did not test for certain
effects, he or she should still be deriving correct knowledge from the conducted experiments. In other words, the outcome of the inductive learning process should be consistent with the overt behavior of the learner.

Per variable, we assessed whether the post-test conclusion correctly states the effect of the variable given the conducted experiments. For each variable, one out of three possible judgments was given:

1. Legal derivation (LD, +): The learner had enough evidence to derive the conclusion stated in the post-test based on a VOTAT strategy and there is no evidence in the experiments that the derivation might be incorrect.
2. Confirmation bias (CB): The reported conclusion is inconsistent with the conducted experiments. That is, there is a set of experiments on the basis of which a different conclusion should have been derived.
3. Ungrounded conclusion (UC): The conclusion the learner gives is not reflected in the conducted experiments; there is no set of experiments that can be used to infer knowledge about the effect of the levels of this variable.

Based on the above definitions, all conclusions in tasks with discontinuous levels of the variables can be categorized. The scoring procedure is as follows. Each judgment is based on the available set of experiments, defined as discussed above. If there is a set of experiments that makes a legal derivation of the effect mentioned in the post-test possible, assign LD to that variable. If in a later set of experiments the originally found effect is shown to be incorrect, the learner can either change the effect of the variable or decide to stick to the old, now “rejected” effect. If the learner's behavior reflects the latter, assign CB to that variable. Note that this limits the assignment of CB to variables that are involved in more complex relations (i.e., interactions). This is therefore a relative limited interpretation\(^2\) of confirmation biases. If the reported effect in the post-test is not reflected in any set of experiments, UC is assigned to that variable.

The problem with these measures is that there is a possibility that the confirmation bias score is too high if a learner is conducting a lot of experiments. That is, after a (large) number of experiments, there is a reasonable chance that there are two

\(^2\)Obviously, only if compared with think-aloud data, this limited confirmation bias can be extended to cover more detailed accounts of confirmation bias (e.g., Wason, 1960; Klayman & Ha, 1987).
experiments that together show the effect of a particular variable. However, if these experiments are separated in time, chances are that the learner never linked these two experiments together. In those situations, the answer given during the post-test might incorrectly be scored as “confirmation bias”. (Because the conclusion given as answer during the post-test is shown to be incorrect in the overlooked combination of experiments.) Therefore, we measured the quality of behavior by comparing per conducted experiment the last experiments with the experiments that were on the screen and with the experiments that the learner scrolled back into view during the construction of the last experiment (operationalized by requiring the user to scroll back to an earlier set of experiments and look at these experiments for at least 2 seconds).

Table 5.2: Peter-task results: Consistency measures per variable, the proportion of experiment-space searched and the comprehension score per learner.

<table>
<thead>
<tr>
<th>Learner</th>
<th>Quality of Behavior per Variable</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bike</td>
<td>B'fast</td>
</tr>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td>CB</td>
</tr>
<tr>
<td>4</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>5</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>6</td>
<td>CB</td>
<td>CB</td>
</tr>
<tr>
<td>7</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>8</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>9</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>10</td>
<td>+</td>
<td>CB</td>
</tr>
<tr>
<td>11</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>12</td>
<td>UC</td>
<td>UC</td>
</tr>
<tr>
<td>13</td>
<td>UC</td>
<td>UC</td>
</tr>
<tr>
<td>14</td>
<td>UC</td>
<td>+</td>
</tr>
<tr>
<td>15</td>
<td>UC</td>
<td>+</td>
</tr>
</tbody>
</table>

Mean: 4.1 0.20 11.3

Note: + = legal derivation, UC = ungrounded conclusion, CB = confirmation bias. B'fast = breakfast, Consis = consistency score, ESpace = proportion of experiment-space searched, Compr = comprehension score.

The scores for this consistency measure are shown in Table 5.2. Although the Pear-
son Correlation between the consistency and the comprehension scores is quite high ($r = 0.79$, $p < 0.01$), there are some notable differences which yield a relative low correlation if the rank orders are compared (Spearman's $r = 0.47$, $p < 0.1$). These differences will be addressed based on examples\(^3\) from Table 5.2.

- Learner 1 scored 12 out of 18 points on the comprehension score measure. According to the notion that the “higher the comprehension score, the better the performance”, this particular score reflects suboptimal discovery behavior. However, according to the consistency measure, this learner scores 5 out of 5. With respect to the number of experiments, this learner conducted exactly 7 experiments ($7/48 = 0.15$), which is the necessary number of experiments to derive all main level effects. Although one might argue that this learner performed suboptimal, and that another approach or the use of other methods might have been better, this behavior is perfectly rational as the learner discovered all main effects with a minimal amount of effort. (Similar reasoning holds for learners 2 and 11. Learners 5, 7, and 9 can be similarly categorized, as they also reported all (main) effects that they observed. Although they conducted more experiments, these experiments did not reveal new information.)

- Learner 3 scored also suboptimal on the comprehension score measure compared to the maximum scores, but higher than the score related to only discovering the main effects (an empirical score of 13 versus a score of 12 if only the main effects would have been reported). Moreover, in the current set of learners, only one learner scored better. Therefore, one might conclude that this learner performed quite well. However, on the consistency score, this learner scored worse (a score of 4 versus the maximum of 5). This is caused by the learner not reporting that the levels of the breakfast variable have an effect although this learner constructed a set of experiments which allowed him to infer this. Similar observations, that is, having a higher comprehension score that can be accounted for on the basis of the consistency score can be made for learners 6, 10, and 14.

- Learner 13 did not engage in organized experimenting (none of the experiments

\(^3\) Note that learner 8 has the maximum consistency score, but a below average comprehension score. This is due to the learner reporting during the post-test that he forgot testing for the effect of the bags, therefore gaining a “+” consistency score but not points for the comprehension score.
was, if compared to the immediate preceding experiment, in accordance with the VOTAT strategy). Even more, the conclusions reported during the post-test were not in line with the evidence. Thus, the learner scored 0 out of 5 points for the consistency measure. However, the learner guessed some of the relations correctly and therefore scored above minimum (4) for the comprehension score. Although not as extreme, learner 12 also showed minimal VOTAT behavior, yielding many uninterpretable experiment. As learner 13, this learner showed a much higher comprehension score than can be warranted on the basis of his unsystematic behavior.

Taken together, if one wants to use the comprehension score as measure for the quality of discovery learning, one neglects the issues raised above. Moreover, it is not clear on what grounds a learner comes to a certain score on this task. It might be due to an extensive battery of discovery skills and methods that are available but which are used in a rather unskillful manner, or a learner might posses little knowledge but use the available resources very efficiently.

**Towards a Consistent Consistency Measure**

As we have argued, one of the problems of the comprehension measure is that prior knowledge influences the score. Therefore, comprehension scores are likely to be influenced by both over-task stable discovery skills and by domain-specific knowledge, yielding an unstable measure when comparing discovery skills over tasks. On the other hand, as was argued earlier, the consistency of a learner’s discovery behavior is not influenced by domain related knowledge. Therefore, the intra-subject correlation between the consistency measures on two tasks in different domains should be higher than the correlation between the two comprehension measures.

We therefore tested the same learners directly after the Peter-task in a task with a comparable task structure, but in a different domain setting. Instead of a boy cycling to school, the learners were told a cover story about a shop in which different types of products (i.e., dairy: milk and butter-milk, fruit: apples and pears, soft drinks: brown, green, and yellow, vegetables: leek and onions, bread: white and whole-grain) could be bought. As dependent variable the total price of the five selected products was given. Moreover, the learners were told that two products together were on discount (i.e., if pears of the fruit product group were bought in combination
with the whole grain bread of the bread product group, the total price was reduced with 20 cents). This statement served as an interaction-clue to ensure that the prior knowledge for learners differed between both tasks. The shopping-task consisted of the same underlying relations as the earlier discussed Peter-task. That is, two variables interact (the effect of two variables is conditional on the level of the other variable), one variable showed a main effect, and two variables did not have an effect. The same measures as for the Peter-task were gathered, and are shown in Table 5.3.

Table 5.3: Shopping-task results: Consistency measures per variable, the consistency score, the proportion of experiment-space searched and the comprehension score per learner.

<table>
<thead>
<tr>
<th>Learner</th>
<th>Fruit</th>
<th>Quality of Behavior per Variable</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dairy</td>
<td>Soft Drinks</td>
</tr>
<tr>
<td>1</td>
<td>CB</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td>UC</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>5</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>6</td>
<td>+</td>
<td>UC</td>
<td>UC</td>
</tr>
<tr>
<td>7</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>8</td>
<td>CB</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>9</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>10</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>11</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>12</td>
<td>CB</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>13</td>
<td>CB</td>
<td>UC</td>
<td>+</td>
</tr>
<tr>
<td>14</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>15</td>
<td>CB</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Mean: 4.1 0.29 12.7

Note: + = legal derivation, UC = ungrounded conclusion, CB = confirmation bias. Consis = consistency score, ESpace = proportion of experiment-space searched, Compr = comprehension score.

As can be seen by comparing Tables 5.2 and 5.3, the combination of the shopping domain and the interaction-clue in the instruction leads to a more extensive coverage of the experiment space (an increase of .09, which equals 45%, $t=-3.73$, $df=14$, $p < .01$, with a paired correlation between domains of 0.56, $p < 0.05$, Spearman $r = .52$, $p < 0.05$), and to a higher comprehension score (an increase of 1.4, which equals 12%, $t=-2.59$, $df=14$, $p < .05$, with a correlation of .48, $p < 0.1$, Spearman
Measures of Inductive Learning Quality

$r = .41, p > .1)$. One might argue for both the number of unique experiments and the comprehension score that an increase for the shopping task is not surprising as, in contrast to the cycling task, the learners are being told about the possible relations beforehand.

However, if one would consider the comprehension score as a valid and stable measure of the quality of domain- and probably task-independent discovery skills, the comprehension scores on both tasks should have a higher correlation than observed. Even more, a valid test should show a consistent test-retest pattern. Therefore, the mean difference between the two comprehension scores should be close to zero. However, the t-test showed that there is a significant difference between the comprehension scores for the shopping and the cycling variants of the task.

Obviously, this difference might be attributable to the additional instruction that the learners got in the shopping task. However, this merely has the effect that learners have more knowledge about the underlying properties of the task in one setting compared to another. Given that it is unlike that that influences the quality of the discovery skills, this difference shows the inappropriateness of using the comprehension score as measure of those skills.

In contrast to the difference between the comprehension scores on both tasks, the consistency score is more coherent between the two tasks. The mean consistency score for both tasks is 4.1, and the correlation between the scores on the Peter-task and the shopping-task is .72 ($p < 0.01$, Spearman $r = .50, p < 0.05$). Because of the (almost complete) absence of a mean-difference, a t-test does not allow the rejection of the $H_0$ hypothesis. Power analysis showed that given the current experimental setup and data-set, a true difference larger than 0.6 between the means would have been detected with a $\beta$ of .05. Combined with the results presented above, this warrants the conclusions that the individual consistency scores are more alike than those for the comprehension score.

Although more extensive studies need to be conducted to more thoroughly test this new consistency measure, the studies presented above show that a different aspect of the discovery behavior is assessed, that this measure is less influenced by differences in prior knowledge and that the consistency score is relatively consistent with respect to intra-subject performance.
CONCLUSIONS

We proposed a new measure for the quality of discovery behavior, based on an analysis of discovery behavior. This measure expresses the quality of the discovery process based on the consistency of the discovery process compared to the answers given during the post-test instead of the completeness of the discovery process as expressed in the coverage and comprehension scores. The newly proposed consistency score measures the quality of the evidence interpretation, an important part of the scientific discovery process as described in the SDDS-related theories. In contrast, the comprehension score cannot be related to a particular part of the scientific discovery process. The quantitative measures used in previous research focused on the number of relations found or the proportion of the experiment space covered by the experiments. However, as argued, the goal of the inductive discovery task is inherently ambiguous. If the learner has not been given specific information about when to stop, it is impossible for the learner, to know when all necessary knowledge has been collected, as the next experiment can always reveal a new, previously unknown relation. The stop-criterion is therefore always subjective and depends on the expected usefulness of testing more hypotheses or conducting additional experiments. If the costs of continuing becomes too high compared to the probability of finding better or more relations, it is rational to stop experimenting. And as the expected usefulness of a certain hypothesis is related to the individual's prior experience, one cannot judge the lack of testing certain hypotheses as not testing might be the most rational option available.

Nevertheless, the comprehension score measure remains useful. Not as a pure measure of the quality of the discovery behavior, but as a measure to what extent the learner was able to discover the relations in the domain given his prior knowledge and discovery skills. Therefore, this measure can be successfully applied to compare behavior (within subject) on similar domains, for example to study transfer or training. Moreover, the comprehension score gives more information about the completeness of the discovery process than the consistency score. That is, the consistency score is not weighted for the completeness of the answers during the post-test: a learner can score high on the consistency score without discovering most relations. For example, a learner in the Peter-task can achieve a maximum score on the consistency measure without testing for interactions, whereas a learner who does test for
interactions might score lower on the consistency score if this learner is not able to derive the interactions correctly. Therefore, to assess the potential discovery behavior of a learner, one should analyze the learner's behavior to see which strategies are used. But, as argued before, this needs to be done in a number of different tasks or domains such that accidental discoveries and bad luck average out.

But with respect to measuring the quality of the inductive learning process, the comprehension score measure is not appropriate. Learners who accidentally guess a relation are credited, although their correct answer is not related to discovery behavior. Similarly, if learners discover that one of their earlier discovered relations is incorrect, but do not refine this relation, their comprehension score remains the same. Not investigating that yet unknown relation is at least suboptimal discovery behavior.

To overcome this problem, a quality score should ideally reflect the above sketched deviations from optimal discovery behavior. In other words, a quality score should reflect only those relations which are correctly discovered. To judge whether a relation is discovered correctly, the behavior of the learner during the task has to be examined. Based on the set of experiments that the learner conducted, one can derive which relations mentioned during the post-test are correctly derived and which relations are either not valid or incomplete based on the learners dataset. Although this measure does not indicate which strategies or what prior knowledge is used by learners during the discovery process, the quality in the sense of correct usage of discovery skills can be objectively assessed, based on relative straightforward log-file analyses. Therefore, the use of a comprehension score should always be accompanied with a measure reflecting the consistency of the behavior.
APPENDIX 4.A: THE NECESSARY NUMBER OF EXPERIMENTS

An important concept in the analysis and evaluation of the inductive learning process is the number of experiments that a learner conducts. If a learner has reasons to expect that in a domain only main effects can occur, it is not sensible to test for first or higher order interactions. On the other hand, if interactions can be expected, doing only those experiments necessary to find the main effects is not sensible either. Therefore, the most sensible number of experiments is dependent on the expectations a learner has about the domain.

To find all effects in a static domain (i.e., a domain without non-deterministic and sequential effects), one has to do all possible experiments, that is, all combinations of all levels of all variables. This can be calculated by multiplying the number of levels of all variables. For example, in the Peter-tasks, this leads to $2^4 \times 3 = 48$ experiments. However, determining the minimum number of experiments to discover the main effects, or the number of experiments to discovery second order interactions is less straightforward. In the sections below, a number of formulas are presented to calculate the necessary number of experiments to find specific effects. The following naming conventions are used in these formulas:

\[
N_{(\text{effect})} \quad \text{The number of experiments needed for the discovery of "effect".}
\]

\[
V \quad \text{The total number of variables.}
\]

\[
\text{lev}(v_n) \quad \text{The number of levels of variable } n.
\]

Table 5.4: Naming convention for formulas 5.1 to 5.3.

MAIN ORDER EFFECTS

To discover a main effect of a two level variable given a context with other variables, the learner has to compare two experiments in which all but that variable have the same level for both experiments. Only given these experiments, one can attribute the found effect to the changed level (Tschirgi, 1980; Schauble, 1996; Chen & Klahr, 1999). For a three level variable, three of these type of experiments are necessary.

Although a learner needs to compare two experiments for a two level variable, and three experiments for a three level variable, the number of experiments necessary to discover all main effects is not the sum of all levels for all variables. That is,
the learner can use the initial experiment as a base experiment to compare other experiments to. Therefore, a learner needs a base experiment, and one additional experiment for each pair of levels per variable. For example, in Table 5.5 one can see that only seven experiments are needed to find the main effects in the Peter-task.

<table>
<thead>
<tr>
<th>1 2 3 4 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 A A A A A</td>
</tr>
<tr>
<td>2 B A A A A</td>
</tr>
<tr>
<td>3 A B A A A</td>
</tr>
<tr>
<td>4 A A B A A</td>
</tr>
<tr>
<td>5 A A C A A</td>
</tr>
<tr>
<td>6 A A A B A</td>
</tr>
<tr>
<td>7 A A A A B</td>
</tr>
</tbody>
</table>

Table 5.5: Seven Experiments to Find the Main Effects in the Peter Task.

More formally, this leads to the following formula:

\[
N_{(\text{Main})} = 1 + \sum_{i=1}^{V} (\text{lev}(v_i) - 1)
\]  

(5.1)

**First Order Interactions**

To detect first order interactions, the learner first has to deduce that an effect is not caused solely by a main effect. Therefore, the same experiments that were needed to discover the main effects have to be conducted \(N_{(\text{Main})}\). After these experiments, the learner should conduct those experiments that are lacking from the set of “all combinations of the levels of pairs of variables”. For example, if a learner wants to know the number of experiments necessary to find first order interaction between two three level variables, 4 experiments should be conducted in addition to the 5 experiments necessary of the main effects. This can be seen in table 5.6. The first 5 experiments are for identification of the main effects, while experiments 6 to 9 are necessary for the other combinations. To find the interactions, each level of the first variable has to be combined with each level of the second variable. However, because in the first 5 experiments all levels of the second variable were already combined with the first level of the first variable, only two \((\text{lev}_{v1} - 1)\) levels remain to be tested. Similarly, the second and third level of the first variable are already tested with the
first level of the second variable. Therefore, the number of extra experiments needed is \((\text{lev}_v - 1) \times (\text{lev}_v - 1)\).

\[
\begin{array}{cccc}
1 & 2 & 3 & 4 \\
1 & A & A & \\
2 & B & A & \\
3 & C & A & \\
4 & A & B & \\
5 & A & C & \\
6 & B & B & \\
7 & B & C & \\
8 & C & B & \\
9 & C & C & \\
\end{array}
\]

Table 5.6: Experiments necessary for discovering first order interactions between two three level variables.

If instead of two variables the domain consists of three variables, there are additional interactions possible. Instead of the one interaction between variables 1 & 2, there are also interactions possible between variables 1 & 3 and 2 & 3. The number of experiments needed for these extra interaction can be calculated in the same manner, \((\text{lev}_v - 1) \times (\text{lev}_v - 1)\) for the interaction between variable 1 & 2, and \((\text{lev}_v - 1) \times (\text{lev}_v - 1)\) for the other interaction. Generalizing this reasoning to a variable number of variables leads to the following formula:

\[
N_{(1st\,Order)} = N_{(Main)} + \sum_{i=1}^{V-1} \left( (\text{lev}(v_i) - 1) \times \sum_{j=i+1}^{V} (\text{lev}(v_j) - 1) \right) \quad (5.2)
\]

**SECOND AND HIGHER ORDER INTERACTIONS**

The above reasoning also applies to for the second order interactions. That is, the main effects and first order interaction experiments have to be tested first. Because one combination of levels of each variable pair was already conducted, the number of new experiments is again based on \(\text{lev}_v - 1\), resulting in the following formula:

\[
N_{(2nd\,Order)} = N_{(1st\,Order)} + \sum_{i=1}^{V-2} \left( (\text{lev}(v_i) - 1) \times \sum_{j=i+1}^{V-1} (\text{lev}(v_j) - 1) \times \sum_{k=i+2}^{V} (\text{lev}(v_k) - 1) \right) \quad (5.3)
\]
As can be seen if formula 5.2 and 5.3 are compared, one term is added and the ranges of the summations are modified. To arrive at the formula for the number of experiments necessary for discovering third order interactions, one has to add another of these terms and modify the ranges according to the modifications between formula 5.2 and 5.3. New formulas for higher order interactions can be built in the same way, simply by adding these terms, and modifying the ranges.

**NECESSARY NUMBER OF EXPERIMENTS IN THE PETER-TASK**

As shown in Table 5.7, seven experiments are necessary for testing the main effects. By applying the formulas 2 and 3, the number of experiments necessary for testing the first-order interactions and second-order interactions are easily derived. As the Peter-task has five variables, a fourth-order interaction is the highest possible interaction. The number of necessary experiments are presented in Table 5.7.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Number of necessary experiments</th>
<th>Number of total experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main order effect</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>First order interaction</td>
<td>+14</td>
<td>21</td>
</tr>
<tr>
<td>Second order interaction</td>
<td>+16</td>
<td>37</td>
</tr>
<tr>
<td>Third order interaction</td>
<td>+9</td>
<td>46</td>
</tr>
<tr>
<td>Fourth order interaction</td>
<td>+2</td>
<td>48</td>
</tr>
</tbody>
</table>

Table 5.7: The number of additional experiments that are necessary to test for main order or interaction effects in the Peter-task.
CHAPTER 5.

DISCOVERY SKILLS IN A COMPLEX TASK

Explaining the limited effects of discovery learning in the Optics domain

Abstract

In this chapter, explanations are offered for the often observed meager results of discovery learning in complex domains. Hereto, we introduce a qualitative-reasoning based method to inspect and analyze learners' behavior during discovery learning in Optics. Optics is a graphical real-time simulation environment of geometrical optics. Studies using Optics (Hulshof, 2001; Prins, 2002) found poor learning effects. We propose alternative explanations for the meager results that are based on qualitative-reasoning guided analyses. These explanations focus on the (1) friction between the knowledge tests and the learned knowledge itself, and (2) an underestimation of the complexity of open or unconstraint discovery learning tasks. We argue that the discovery learning skills of learners tested in complex domains like Optics are easily underestimated.

The tasks presented in the previous two chapters of this thesis were specifically designed to study human discovery learning. In a sense, these tasks mimic the classical approach of psychology to strip as many complexities as possible from a certain task so that the behavior induced can be more readily interpreted. However, as findings based on studying human list memory do not necessarily extend to everyday's memory usage, similar issues might arise when studying scientific discovery learning in a
simplified task. One of the obvious differences between real-world discovery learning and tasks similar to the Peter-task is that in a real-world setting the relevant variables are often not known beforehand. Similarly, it is often difficult in real-world discovery learning to construct the kind of perfectly replicable, discontinuous experiments as is standard in most of the empirical studies. Therefore, Optics – a task discussed in the next section – was developed that is more akin to the complex discovery learning as encountered in the wild.

As with hindsight might have been expected, studies using this complex domain (Hulshof, 2001; Prins, 2002) showed that performance was relatively low; only when the instruction was very focussed and the experiment space was limited, positive effects of discovery learning could be found (Prins, 2002, Chapter 5 & 6). Based on these results, one can derive one or both of the following two conclusions: the unconstraint task is too difficult for the learners¹ or the learners did not possess sufficient skills or capacity necessary for discovery learning. Note that explanations for the low performance that aim at a too difficult underlying model do not hold as Van Someren and Tabbers (1998) have shown that learners are able to infer the lenses-law when given an appropriate data-set.

In Chapter 3 and 4 it was argued that the meager scores on the Peter-tasks were not necessarily due to low discovery skills, but that these low scores might have simply been an effect of not having observed the phenomena on the basis of which useful derivations could have been made. This was supported by showing that the learners did show increased performance on those post-test questions that were related to the conducted experiments. In turn, this might be taken as a suggestion that the learners’ potential discovery behavior was better than could be induced from the suboptimal scores on the post-test.

This chapter is concerned with this same issue of meager discovery results, but now applied to the more complex Optics task. Therefore, we will first discuss Optics and the recent work on Optics reported in Hulshof’s and Prins’ recent theses (Hulshof, 2001; Prins, 2002) that provide the empirical data analyzed in this thesis. As the complexity of the Optics simulation prevents a direct analysis of the learners’ behavior as was done in Chapter 3 and 4, we will introduce a conceptual model of the Optics

¹But note that even learners who were supposed to be experts in the field of Optics had a hard time discovering all relations (Hulshof, 2001).
domain. By applying this conceptual model to the Optics data, an analysis is presented that is similar to Chapter 3 and 4 and that can be used to assess: “What causes inductive discovery learning to be difficult?” in this more complex task.

As said, before turning to the analysis of the Optics task and learners’ behavior in Optics, we will first discuss Optics and the rationale for studying behavior in a complex discovery learning setting.

**Optics**

Recent years have shown an increased prominence of computer simulations in innovative educational methods. These computer simulations enable students to learn about complex domains by experimentation, something which is often either infeasible, impractical or too expensive without the aid of computer simulation. This self-directed experimentation is assumed to foster learning processes in which emphasis is on the discovery of knowledge. Traditionally, research on this type of learning is done by studying discovery learning in either its natural contexts (e.g., when the studied task is part of a school curriculum), or in a much more constrained setting (e.g., when the studied task is especially designed for empirical examination). The Optics environment, discussed in this chapter, was developed to fill the gap between these extremes.

Optics is a computer simulation environment\(^2\) of geometrical optics. Care was taken during the development of this simulation environment to keep it as “neutral” as possible. That is, although a simulation will always be a simplification of the simulated domain the simulation was kept as close as possible to a real-world optics setting. For example, although it has been shown that certain task-independent artifacts improve discovery learning (e.g., instructional support, Njoo & De Jong, 1993, or hypothesis scratch pads, Van Joolingen, 1993) these were not included in the simulation, nor were task-specific artifacts added that might guide the learner in the discovery process (e.g., increasing the visual saliency of critical concepts such as the (virtual) focal point). The main difference between a physics lab and the present Optics simulation is that the light-rays are depicted as colored lines in the simulation, whereas inducing the path of the light-ray is one of the tasks a learner has to perform in a physics lab.

\(^2\)Optics was developed using XPCL/SW1-Prolog by Jan Wielemaker (Social Science Informatics, University of Amsterdam).
lab setting. Obviously, artifacts like measuring instruments (e.g., for measuring the
distance between two objects or the angle of light beams) are available to learners in
this simulation.

Figure 6.1 presents a workbench situation as it could have been constructed during
experimentation in Optics. At the start of a new session, the learner is presented
with an empty optical workbench. By first selecting an entity (e.g., a lamp or lens)
and then clicking on the workbench, a learner can add objects to the simulation.
Using the buttons depicting a box with arrows pointing outwards (third and fourth
button from the left on the second toolbar), the learner can reposition the lamp or
the lens. Using the button with a tilted lamp and arrows, the learner can change the
angle by which the light beam leaves the lamp. While dragging the lamp or lens,
the environment immediately reflects the effects of these movements, and the same
holds for the angle by which the light beam leaves the lamp.
The other buttons are involved with conducting measurements (e.g., learners can add construction lines to the environment and measure the distance between previously added construction lines using the sixth and seventh button from left on the first row), requesting help or rereading the assignment (the last three buttons from the right on the second row) or deleting previously added objects (the PacMan-like buttons). A complete description of the Optics implementation can be found in Chapter 3 and 4 of Hulshof’s thesis (2001).

The direct-manipulation of lens, lamp and light-beam highlights the two aspects of the continuous nature of Optics. Not only are the levels of most of the variables (position of the lamp, angle of the light beam) continuous, the construction of experiments also has a continuous nature. This continuity contrasts Optics with most other discovery tasks. For example, in discovery settings like the Peter-tasks (Wilhelm, 2001, Chapter 3 and 4 of this thesis) or the BigTrak (Klahr & Dunbar, 1988) studies, the discontinuous nature of the simulation is easily inferred as both the dependent and independent variables are discontinuous and the experiments generated by the learners are necessarily discrete. But also tasks such as Bubbles (Hulshof, 2001) and HeatLab (Veenman, Elshout, & Meijer, 1997) which do have some continuous elements (i.e., the formulas underlying the systems’ behavior are of a continuous character) are in essence discontinuous as learners can only construct discrete experiments.

For the analyses presented in this Chapter, we will focus on the action-logs and the answers to pre- and post-tests. Each action initiated by a learner is reflected in log files. These log files contain sufficient information to enable the accurate replay of the user’s interactions with the environment. The Optics environment also consisted of a test module which presented the learners with questions and a set of multiple choice answers. Although different types of questionnaires have been used, the most typical questions are shown in Figure 6.2. As can be seen in this Figure, the questions are at a conceptual level instead of numerical or algorithmic level as often used in secondary schools’ curricula. These questions follow the format described as “What-if” questions by Swaak and De Jong (Swaak & de Jong, 2001). Instead of focusing on factual knowledge, these questions are designed to assess “intuitive knowledge” that learners have gained by working in a discovery task. These “what-if” questions present the learners with a depiction of a configuration with an associated question.
Figure 6.2. Four illustrative examples of diagrams and questions used in assessing the prior and post-experiment Optics domain knowledge.

stated in relatively simple terms so that the learner can supposedly answer the question without having to resort to formal reasoning.

EMPIRICAL RESULTS WITH OPTICS

The studies with Optics conducted by Prins (2002) and Hulshof (2001) were aimed at studying the effects of subject-oriented variables (e.g., meta-cognition and domain-neutral and domain-specific prior knowledge) on discovery learning. To keep these studies as close to “natural discovery learning” as possible, no detailed explicit guidance was given to the learner. For example, although all variables that the learner can manipulate were discussed during the pre-experimental instruction, it was up to the learner to identify which of these variables are causally related to the observations that could be made. Moreover, in most of these studies, it was not specified which
variable was to be considered as dependent or independent variable.

Given this set-up, learners in Prins' and Hulshof's studies showed only a small learning effect, even after working for 20 minutes to up to an hour in the most simple configurations (e.g., a single lens and a simple laser-like lamp of which the light beam can be tilted but of which the y-position is fixed on the optical axis). This limited learning effect has been shown with learners sampled from different populations and with different Optics configurations and different instructions. Table 6.1 summarizes the most important results. All results in this Table are based on the most simple configurations in the experiments, typically consisting of a single lens and a single laser-like lamp. The pre- and post-test columns show the percentage of correct answers on the pre- and post-tests. The “Max score” column denotes the number of questions that could be answered correctly based on experience in the most simple phase. The last column is the learning gain, expressed as the difference in percentage correct answered questions in pre- versus post-test. For this assessment of domain knowledge, both the pre- and a post-test contained questions as shown in Figure 6.2.

Table 6.1: Mean percentage correct scores for the (first phases of) five Optics studies. See text for a description of the columns.

<table>
<thead>
<tr>
<th>Study</th>
<th>Pre-test</th>
<th>Post-test</th>
<th>Max Score</th>
<th>Δ%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hulshof, Chapter 5</td>
<td>38%</td>
<td>44%</td>
<td>18</td>
<td>6%</td>
</tr>
<tr>
<td>Hulshof, Chapter 6</td>
<td>54%</td>
<td>47%</td>
<td>24</td>
<td>-7%</td>
</tr>
<tr>
<td>Hulshof, Chapter 7, overall</td>
<td>51%</td>
<td>50%</td>
<td>30</td>
<td>-1%</td>
</tr>
<tr>
<td>no-support condition</td>
<td>52%</td>
<td>46%</td>
<td>30</td>
<td>-6%</td>
</tr>
<tr>
<td>support condition</td>
<td>48%</td>
<td>54%</td>
<td>30</td>
<td>6%</td>
</tr>
<tr>
<td>Prins, Chapter 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qualitative test</td>
<td>16%</td>
<td>22%</td>
<td>36</td>
<td>6%</td>
</tr>
<tr>
<td>quantitative test</td>
<td>13%</td>
<td>17%</td>
<td>12</td>
<td>4%</td>
</tr>
<tr>
<td>Prins, Chapter 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qualitative test</td>
<td>35%</td>
<td>51%</td>
<td>15</td>
<td>16%</td>
</tr>
<tr>
<td>Prins, Chapter 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no-support</td>
<td>40%</td>
<td>62%</td>
<td>12</td>
<td>22%</td>
</tr>
<tr>
<td>support-condition</td>
<td>45%</td>
<td>64%</td>
<td>12</td>
<td>19%</td>
</tr>
</tbody>
</table>

Even considering that the reported studies were not designed to optimize knowledge transfer, the effect of discovery learning was relatively low in the less constrained studies (Hulshof, Chapter 5, 6 & 7, Prins, Chapter 4). Without additional support,
as was implicitly given in Prins’ Chapter 5 and 6 studies by focusing the instructions and constraining the experiment space, learners show at best a very low performance gain. That is, the studies reported by Hulshof and the first study reported by Prins were relatively unconstrained. The learners were not given a specific instruction and even the most simple phases were relatively complex. The two last studies reported by Prins were more constrained, both by giving the learners a more detailed assignment and by limiting the complexity of the environment.

This limited positive effect of discovery learning cannot simply be attributed to the deficiencies in the tested learners’ behavior, as they did show positive learning effects on other discovery tasks. For example, the learners in Hulshof’s Chapter 5 experiment also participated in the Bubbles task in which they showed improved performance (see Hulshof, 2001, paragraph 5.3.1.3 and Table 5-7). Similarly, it has been shown that learners sampled from the same population as Prins’ studies are able to discover a law similar to the lenses law when they are presented with an appropriate data set that contains observations for the three important variables (Van Someren & Tabbers, 1998).

**Assessing Discovery Skills**

A warranted question after studying the meager results is what this means for the viability of discovery learning in the Optics task. Are the populations from which Hulshof and Prins sampled the participants for their studies incapable of self-directed discovery learning in this specific setting? One might be tempted to support this hypothesis as participants answered only approximately two more questions correct after experimenting in Optics. However, this question can only be answered when the learners’ behavior is taken into account. As has been shown for the Peter-tasks, low learning gains are not necessarily indicative of bad discovery skills. It might be that the learners possess the appropriate discovery skills but that they, for example because of a lack of knowledge about what to look for, are not able to utilize their skills to the fullest extent. For the Peter-task, this notion was supported by comparing the conducted experiments with their post-test answers.

The most striking difference between the Optics task and the Peter-tasks is the much higher level of discreteness in the latter. In the Peter-task, a relative direct assessment is possible of the possibly derived knowledge by analyzing the experiments the
learner has constructed (and if necessary aligning think-aloud protocols with those experiments). However, in a task like Optics, experimentation is most of the time continuous. When a learner moves a lamp on the x- or y-axis, a theoretically infinite number of experiments could have been observed. Obviously, a learner does not perceive all these experiments, but there is no straightforward way of inferring what situations during that manipulation are perceived by the learner. Therefore, an different approach is necessary to compare learners’ Optics behavior with post-test results.

The approach taken is to construct a qualitative model of the Optics task. In a qualitative model (Forbus, 1984), one specifies which entities of a system are important in explaining the behavior of the system. A computational version of a qualitative model has the advantage that one can easily test whether the selected and modeled entities are sufficient for explaining the observed behavior and whether they are internally consistent. Moreover, it could eventually be integrated in the Optics simulation to enhance the learning experience. Another reason why developing a qualitative model is relevant, is that the post-test questions are also at a qualitative level. Using a qualitative model provides information on what concepts underly successful answering of a certain question.

For studying the discovery behavior of learners in Optics, we focused on the first phase of the study presented in Prins’ Chapter 5. In this phase, the learners can manipulate the horizontal positions of lamp and lens. The qualitative model should be able to qualitatively represent the causality of the learners’ manipulations. That is, the model has to derive and describe the changes in the system given a modification of a variable. Given such a model, we can (1) identify qualitatively different states in the simulated domain, (2) use the combination of log files and the identified states to assess what situations a learner has observed, (3) construct an overview of the knowledge the learner might have inferred from the observed states and (4) match the inferences against the knowledge a learner reports in the think-aloud protocols and during the post-test. By identifying which states have been seen by learners, we

---

3 Each horizontal manipulations of the lens can also be expressed as a movement of the lamp. For example, moving the lens one centimeter to the left is analogous to moving the lamp one centimeter to the right. Therefore, all horizontal movements are expressed in the remaining of this chapter as manipulations of the lamp. Although this could be one of the issues a learner might focus on during experimentation, none of the learners made references to this in the analyzed think-aloud protocols.
can assess the quality of the learners' discovery learning using a similar method as done for the discontinuous Peter-tasks. Moreover, we can assess the complexity of a domain by comparing the number and type of states between tasks.

**THE QUALITATIVE MODEL.**

The qualitative model reported in this chapter was developed in the GARP architecture\(^4\) (for a more thorough description of GARP, see Bredeweg (1992) or De Koning, Bredeweg, Breuker, and Wielinga (2000)). Simulations developed in GARP are best thought of as a knowledge based description of behavior. A GARP simulation starts with an initial scenario and generates a graph of all possible qualitative distinct states for the domain described in the scenario. New states are generated by applying model fragments and transition rules. Transition rules describe what the consequences are of ongoing changes. For example, if a certain value is decreasing, a transition rule will generate a new state that reflects that the value has to reach a state that reflects a natural "0". These rules are of less importance in the current model, so will not be discussed in detail. Model fragments consists of a set of conditions and consequences. If the conditions are met, the current state is changed to reflect the consequences referred to in that fragment. Note that new states, whether they are derived from transition rules or applying model fragments, are only added to the graph if the new state differs qualitatively from the previous state. In the Optics domain, an example of a state change is when the x-position of the lamp is changed from "left of the virtual focal point" to "on the virtual focal point".

The values of quantities, like "x-location" in the example above, are expressed on a quantity space. A quantity space represents the variable aspects of an entity. Instead of being represented in actual numeric values, the values of the quantities are expressed on an ordinal scale. Each value on this ordinal scale represents a qualitatively different situation. For example, the "x-location" of the lamp can be either to the left of a virtual focal point (LoVFP), on a virtual focal point (VFP), or to the right of a virtual focal point (RoVFP), making a quantity space that can be expressed as \(<\text{LoVFP}>\), \(<\text{VFP}>\), \([\text{RoVFP}>\). In this notation, \(<\text{VFP}>\) represents a single point, whereas \(<\text{LoVFP}>\) and \([\text{RoVFP}>\) represent intervals bounded at one side by the

\(^4\)In comparison with ACT-R, the modeling architecture used for the model in Chapter 2, GARP is not focussed on explaining human cognitive behavior per se, but can better be thought of as a reasoning engine.
virtual focal point (respectively right and left). These points, called landmarks, are defined to capture all qualitative distinctive behaviors in the target system. That is, even if the virtual focal point is not visible to a learner, it does determine the target system (i.e., Optics) behavior. In other words, in each of the different quantity space levels the qualitative behavior of the modeled system is different. At the same time, different quantitative settings within one state yield the same qualitative behavior. Besides expressing the value of quantities, the quantity spaces are also used to express the derivative of the value. To express that the quantitative value of that quantity is increasing (or decreasing or stable), the derivative of that quantity is set to plus (or minus or zero).

As the Optics simulation environment does not include any self-changing properties (i.e., like water flowing out of a container, which will eventually yield a new state representing an empty container), all changes in states are human induced. These actions are represented as agent model-fragments. This is an important concept in the Optics model as these agents represent the possible modifications in the simulation invoked by learner actions. The two changes invoked by agents are changes in the lamp position and changes in the angle of the light beam.

The operations initiated by these agents can change the quantity value of different entities. For example, when the lamp is moved from $<\text{LoVFP}>$ to $[\text{VFP}]$, the model-fragment representing the intersection between the light beam at the right side of the lens and the optical axis changes to reflect the lack of intersection. This influence of one quantity on another is an example of a dependency. Dependencies relate different quantities, either within or between model-fragments. For the purpose of the Optics model, two causal dependencies are of most interest: proportionalities and influences (Forbus, 1984). If two quantities always change concurrently, this is modeled by a proportionality. If a change in a quantity causes a change in another quantity in the same direction, this is expressed as a positive proportionality ($prop-pos$). A negative proportionality ($prop-neg$) causes a change in the opposite direction. An influence expresses that the value of a quantity determines the derivative of another quantity. A positive influence ($inf-pos$) indicates that if a quantity's value increases, the derivative of another quantity is positive (and vice versa). A negative influence ($inf-neg$) relates an increasing quantity with a negative derivative for another quantity (and vice versa).
In the next paragraph we will discuss the GARP building blocks for the qualitative model of the Optics simulation.

**Optics' Entities, Quantities and Quantity Spaces** The simulation of Optics in GARP is centered around three major entities: the lens, the light beam and the lamp. The lamp is the central entity of these three. That is, the horizontal position of the lamp is expressed as a distance from the lens, and the starting position of the light beam is obviously defined by the horizontal position of the lamp. As the vertical position of the lamp is fixed in the experiment simulated with the GARP model, the two variables of interest are the lamp's horizontal position compared to the lens and the angle of the light beam with respect to whether or not it intersects with the lens.

From a qualitative point of view, the horizontal position of the lamp can take one of three values. The lamp can be positioned left of the virtual focal point, on the virtual focal point, and between the virtual focal point and the lens. Although the lamp can also be positioned to the right of the lens, this is not captured in the simulation as (a) this does not lead to new, valuable information and (b) none of the learners actually deliberately constructed this situation. The quantity space of the light beam is also related to the lens, that is, although the actual manipulation of the light beam is in terms of the angle from the lamp, the consequence of a certain angle is best expressed in terms of where the light beam intersects with the lens. (Note that it is not possible to create a qualitative model of Optics without incorporating dependencies between components.) The light beam, emitted from the lamp, can either not hit the lamp by passing over or under the lens, or hit the lens. If the light beam hits the lens, the location of the intersection between light beam and lens can be categorized into one of three states. The light beam can intersect with the lens above the optical axis, below the optical axis, or the light beam can hit the lens in the optical center of the lens\(^5\). These quantities and their quantity spaces are depicted in Figure 6.3. In this Figure, panel A is a schematic version of an Optics configuration. Panel B shows the qualitative structure of this same configuration; lines with arrows on both sides depict intervals (representing multiple quantitative values), circles denote points (representing a single quantitative value). The significance of a point is that it is the only quantitative value in which certain qualitative relations hold. For example, the

\(^5\)As in the experiments of Prins and Hulshof, the GARP simulation assumes ideal lenses.
only situation in which the light beam is parallel to the optical axis is when the lamp is exactly located on the virtual focal point. Therefore, each change of the x-location variable results in a qualitative change, whereas a change in the x-location when the lamp is at the left of the virtual focal point does not necessarily imply the transfer to a new state.

Whereas Figure 6.3 gives a schematic overview in terms of the domain used, Figure 6.4 gives a VISIGARP overview (Bouwer & Bredeweg, 2001) of the entities and quantities as represented in the GARP model. As can be seen, most of the knowledge is directly related to the lamp. The boxes outside the lamp represent the two user actions: Changing the angle of the light beam and changing the horizontal position of the lamp. The arrows denote dependencies.
The Construction of the State Graph  Panel C and D of Figure 6.3 show two different states that can be derived from the state depicted in Panel A. This initial state describes a situation which is defined by the lamp being placed left of the virtual focal point and the light-beam pointing upwards. The angle of the light-beam is such that it is below its maximum angle but above the angle that would result in hitting the lens (in terms of the intersection quantity space, it is “above the lens”).

Initially, the derivative of the x-position is zero; the state is in equilibrium. This reflects that one cannot derive any causal relations if nothing is manipulated (Figure 6.3, Panel B). If an agent model-fragment is applied that modifies the x-position, GARP calculates which new states are possible. Given a modification of the horizontal-position of the lamp, a new state may arise in which the light beam hits the lens (Figure 6.3, Panel C). This is inferred by the application of a transition rule after the application of an agent model-fragment. However, a change in x-position might also yield the state as depicted in Figure 6.3, Panel D. If this transition rule was applied, it signaled a change in the qualitative value of the x-position of the lamp itself. In this state, the lamp is on the exact position of the virtual focal point. Although the
y-position of the intersection was also changed, this change did not result in a new qualitative state as it is still “above the lens”. Therefore, the value of the intersection quantity is still the same.

GARP continues to apply model-fragments and dependencies until it has generated all possible states, and is able to construct a complete graph which also specifies all state-transitions, an example of which is given in Figure 6.5.

**Applying GARP** As the Optics GARP model is a complete model of the qualitative relations in the Optics domain in Prins’ first experiment, the model contains all qualitative knowledge from scratch. However, a learner who is new in the Optics setting presumably does not know about these relations. Therefore, by means of experimenting, the learner has to construct experiments to see the effect of manipulations. These experiments can be traced in the GARP simulation, which reflects what knowledge a learner could have inferred. The technique of describing a learner’s behavior on the basis of a model of the underlying domain is referred to as model tracing (Anderson, 1990; Anderson et al., 1990; Jansweijer, 1988; Jansweijer et al., 1989). We will use a variant of this technique later in this chapter to analyze learner’s behavior in Optics.

Besides a tool for tracing the experiments, a GARP model of Optics can also be used
for analyzing the complexity of the domain and the size of the experiment space of the Optics task. As referred to earlier, one of the challenges of the Optics domain is the continuous character of the Optics simulation. As the GARP model identifies all qualitative states the domain can take, it can be used as a tool to determine the qualitative experiment-space. In the next section, we will use the GARP model for this purpose: analyzing the domain to identify the number of experiment-space states.

ASSESSING THE SIZE OF OPTICS’ EXPERIMENT-SPACE

From a quantitative perspective, the optics task’s experiment space is infinite. As both the x and y positions of the lamp and the angle of the light beam are variables with a continuous scale, the domain has an infinite number of experiment-space states. However, with aid from the GARP model, the infinite experiment space can be categorized in a distinct number of qualitative different regions. Using this qualitative experiment space, the behavior of learners can be categorized according to the domain-related completeness of their experiment-space search.

STATIC EXPERIMENT SPACE

To categorize the quantitative experiment space, the GARP notion of quantity spaces plays an important role. The qualitative landmarks that define the points and intervals of a quantity space lend themselves relatively straightforward for defining the experiment space.

For the first phase of Prins’ experiment, the following qualitative landmarks exists:

Horizontal position of the lamp:

1. left of the virtual focal point
2. on the virtual focal point
3. right of the virtual focal point

Vertical position of the lamp:

1. not applicable, as the lamp is fixed on the horizontal axis

Beam angle:
1. up and over the lens
2. up and in the lens
3. straight forward, through the heart of the lens
4. down and in the lens
5. down and below the lens

As the beam angle can be in any of the five positions irrespective of the horizontal position, this categorization leads to $3 \times 1 \times 5 = 15$ qualitatively different experiment-space states.

However, given these 15 "static" experiment-space states, a learner cannot derive the relations in the Optics domain. In contrast with more static discovery environments like the Peter-task, constructing one experiment in each static experiment-space state is not sufficient to induce all relations. For example, assume that a learner constructs two experiments, one in which the angle of the light beam is set at 0 degrees and one in which the angle of the light beam is set at 45 degrees, entering the lens just below its top, and leaving it with a refraction of 90 degrees. Based on these experiments, the learner could as well infer that refraction is a continuous function as that refraction is constant at 90 degrees. Therefore, one should construct multiple (at least two) experiments within each qualitative state to test the generality of the findings. Doing so allows the learner to discover that the degree of refraction is a continuous function.

Therefore, one of the most notable differences between Optics and most other discovery environments is the necessity and possibility to change the variables continuously. This makes it necessary to extend the number of experiment states as defined above with the manipulation that the learner can conduct. So, a more extensive experiment-space definition is needed that includes transitions between states and manipulations within states.

**A Dynamic Experiment Space of Optics** To be able to determine the number of dynamic experiment-space states, we again refer to the structure devised for the GARP model of Optics. In this model the optical axis (defining the position of the lamp in relation to the lens) and the lens (defining the effect of the angle of the light beam) are represented as sequences of intervals and points, also referred to as landmarks.
For example, the representation of the lens is shown in Figure 6.6. In this figure, two types of intervals are shown. The interval “above the lens” ranges from infinity to just above the lens (denoted by a vertical “\[-\rightarrow\]”) whereas the interval “in the lens” ranges from just in the lens to just above the heart of the lens (denoted by a vertical “\[-\cdot\]”). Besides the intervals, this figure also contains a point: the heart of the lens (denoted by a circle). This differentiation makes it possible to distinguish the number of possible different dynamic experiment-space states:

1. In an interval that is bounded by landmarks on two sides “\([-\cdot\]”\), four different dynamic experiment-space changes are possible. One can change the state in one direction without moving to a new state, or this change can result in a different (static) state. The same modifications are possible if one changes the state in the different direction, resulting in four different dynamic experiment-space states. This will be referred to as a “two-landmark interval”.

2. In an interval that is bounded on one side “\([-\rightarrow\) or \(<\cdot\cdot\]”\), three different dynamic states are possible. Towards the bounded side, one can change the state resulting in either no static experiment-space state or resulting in a static experiment-space state. However, on the unbounded side, it is not possible to change the state resulting in a different static experiment-space state, resulting in three different dynamic experiment-space states. This will be referred to as a “one-landmark interval”.

3. In a point, only two different dynamic states are possible. That is, a change resulting in a different static experiment-space state, or the complementary change in the other other direction. This will be referred to as a “point”.

The number of possible dynamic state changes are depicted in Figure 6.6 by the vertical arrows. Although all these arrows denote a dynamic change, only the ones crossing a marker (“\[-\cdot\]” or “\[-\cdot\]”) represent a change in qualitative state.

This allows for determining the number of dynamic experiment-space states for the Optics domain, as the values of the two independent variables (angle of the light beam and the horizontal position of the lamp) can be described as above.

**Changing the Angle of the Light Beam** Figure 6.7 illustrates the four different dynamic experiment-space states possible for the situation depicted in the center of
Figure 6.6. Schematic overview of possible experiment-space states with respect to the lens.

Figure 6.7. If the light beam is tilted upwards, either a situation as depicted in Subfigure 6.7a or 6.7c results, depending on the size of the change. If the beam is tilted downwards, either the situation in Subfigure 6.7b or 6.7d results. Similar transitions are possible for the other values in the quantity space, yielding the (counting from top to bottom) $3 + 4 + 2 + 4 + 3 = 16$ dynamic experiment-space states depicted in Figure 6.6. These 16 different states are possible regardless of the x-position of the lamp. However, it might be that the effect of these dynamic experiment-space states differs per x-position of the lamp. Therefore, as there are three different qualitative states with respect to the x-position, a total of $3 \times 16 = 48$ different dynamic experiment-space states are possible with respect to the lens if the x-position of the lamp is taken into account.

**Horizontal Movements** The situation for the dynamic experiment-space states related to the horizontal position is somewhat more complicated. That is, whether or not a horizontal movement yields a static experiment-space state transition depends on the angle of the light beam. If the angle is 0, that is, if the light beam is equivalent to the optical axis, moving the lamp does never result in a change of the static experiment-space state. On the other hand, if the beam angle is not zero, seven different situations are possible. (1) The combination of lamp position and angle can result in no intersection between lens and light beam, depicted in Figure 6.8. In this situation, three dynamic experiment-space states are possible. Either the lamp can be
moved away from the lens, not changing the static experiment-space state (Subfigure a) or the lamp can be moved towards the lens which can either result in a different static experiment-space state (Subfigure c) or not (Subfigure b). If there is an intersection between light beam and lens, the position of the lamp with respect to the virtual focal point has to be taken into account. (2) If the lamp is left of the virtual focal point and the lens, four dynamic experiment-space states are possible. The lamp can be moved to the left removing the intersection between beam and lens, or can be moved to the left but the light beam remains in the lens. If the lamp is moved to the right, the lamp might cross the virtual focal point (resulting in a different static experiment space) or might still be before the virtual focal point after the move. (3) If the lens is on the virtual focal point, the lamp can either be moved to the left or to the right of the virtual focal point, resulting in two different static experiment-space states. (4) If the lamp is right of the virtual focal point, the lamp can be moved to the left or right without changing the static experiment-space state, or can be moved to the left resulting in a different static experiment-space state as soon as the virtual focal point is crossed. (5, 6, 7) The scenarios 2, 3 and 4 are also possible if the simulation is mirrored in the optical axis, that is, both for a lens-crossing above the
As can be seen in Table 6.2, the movements of the lamp on the horizontal axis result in $2 + 2 \times (3 + 4 + 2 + 3) = 26$ different dynamic experiment-space states.

Summarizing, although the total number of qualitative distinct states is only 15, the number of transitions the learner has to make to cover all relations is 74 $(48 + 26)$. Note that numbers are based on the most simple Optics configuration used, that is, one in which the learner can only change the horizontal position of the lamp and the beam angle - and excludes the movements of the lens. This clearly shows the relative complexity of the Optics domain compared to domains used in other discovery tasks. That is, a experiment space of 48 states (see Chapter 3 and 4) in the Wilhelm’s (2001) and Kuhn et al.’s (1995) experiments, and (focusing on the number of qualitative states) 6 qualitative different states (see (Klahr & Dunbar, 1988)) for the BigTrak experiments. Moreover, most of these states are not readily observable in Optics, as no visual cues indicate their states’ existence – so one could argue that the complexity is even larger if one also has to discover the landmarks themselves.
Table 6.2: Overview of dynamic experiment-space states for different horizontal positions.

<table>
<thead>
<tr>
<th>Angle</th>
<th>Scenario</th>
<th>Configuration</th>
<th>Landmarks</th>
<th>Dynamic states</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha = 0 )</td>
<td>1</td>
<td>intersection</td>
<td>Point</td>
<td>2</td>
</tr>
<tr>
<td>( \alpha &gt; 0 )</td>
<td>2</td>
<td>no intersection</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>x-pos &lt; VFP</td>
<td>intersection</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>x-pos = VFP</td>
<td>Point</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>x-pos &gt; VFP</td>
<td>intersection</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>( \alpha &lt; 0 )</td>
<td>6</td>
<td>no intersection</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>x-pos &lt; VFP</td>
<td>intersection</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>x-pos = VFP</td>
<td>Point</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>x-pos &gt; VFP</td>
<td>intersection</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

Note: \( \alpha \) is the angle of the light beam compared with the optical axis, x-pos is the horizontal position of the lamp, and VFP is the virtual focal point.

**DERIVING KNOWLEDGE BASED ON A QUALITATIVE ANALYSIS**

To show that the simplification that is necessarily associated with a qualitative model does not hurt the derivational power, we will show that one can derive the lenses law based on an analysis that does not use quantitative knowledge. Given this goal, this should be considered an analysis of the validity of the *qualitative model-based representation of the task* instead of an analysis of the *interaction between learners and the task*. Therefore, it does not take constraints on mathematical knowledge or reasoning skills into account.

The reason for choosing the lenses law as goal for this task analysis is that this law can be used to explain all behavior in the simulated lenses domain. Given that the learners are instructed to discover as much as possible about the domain, their discovery would have to be considered perfect if they manage to discover this formula. The most common form of the lenses formula is:
In this formula, $f$ is the focal point, $d_o$ is the distance between lens and object (i.e., the lamp), and $d_i$ is the distance between the image and the lens. This is depicted in Figure 6.9 in which a lamp and lens are placed on a workbench (which therefore equals the optical axis of lenses placed on it). The dotted lines depict light beams. The light beam starting at $f$ depicts the path of a light beam if it’s originating lamp would be exactly on the focal point. Aside from the distances, this Figure also labels the important angles ($\alpha, \alpha', \beta$ and $\beta'$).

In the experimental setup as used by Prins, the variables that can be changed by the learner are limited to the horizontal position of both the lamp and the lens (defining $d_o$), selecting different lenses (influencing $f$ and $d_i$, unknown to the learner) and the angle ($\alpha$) with which the light beam leaves the lamp. The learner is also told in the initial instruction that special attention has to be paid to the intersection between the optical axis (i.e., workbench) and the light beam (defining $d_i$). In the reasoning below, we will assume that the learner has placed a lamp and a lens on the workbench, with the lamp positioned to the left of the lens.

In the initial situation, the light beam is parallel to the optical axis. As the light beam goes through the heart of the lens, different positions of the lamp do not have any
influence in qualitative terms. That is, regardless of the positioning of the lamp, the light beam will remain parallel to the optical axis. However, changes in $\alpha$ do result in interpretable data. When $\alpha$ is changed, an intersection between the beam and the optical axis appears, triggering the perception of $d_i$. Even more, consequent changes in $\alpha$ without changing $d_o$ show that $d_i$ does not change for changes in $\alpha$. However, this might be an observation that is specific for a combination of a specific $d_o$ value and the currently used lens. But, experimenting with changes in $\alpha$ for different $d_o$ values and for different lenses show that $d_i$ is constant for each combination of $d_o$ and lens. Using Euclidean geometry, one can therefore infer that the relation between $\alpha$ and $\beta$ depends on $d_i$ and $d_o$, and vice versa. However, as none of these properties is known and different lenses result in different $d_i$ values, the only regularities that can be derived from this finding are:

1. Given constant $d_o$ and lens, $\alpha$ is proportional to $\beta$.

The above contains a simplification with respect to the manipulations of the lamp. If the lamp is moved close to the lens (decreasing $d_o$), the point of intersection moves away from the lens (increasing $d_i$). Even more, with the lamp at a certain location, the learner might notice that the light beam runs parallel to the optical axis, and if the lamp is moved to the right of that point, the light beam refracts from the optical axis. In more formal terms, assuming $f$ is taken as a representation of that value of $d_o$ where the light beam is parallel to the optical axis: (1) if $d_o > f$, decreasing $d_o$ increases $d_i$, and vice versa, (2) if $d_o$ and $f$ become equal, $d_i$ approaches infinite, and (3) if $d_o < f$, the refraction is from the optical axis, indicating that there is no intersection between light beam and optical axis. However, in this latter case the light beam is extended backwards, a new intersection appears, for which the same properties hold as on the right side of the lens. Obviously, this virtual intersection is also related to a similar $d_i$ as on the right side of the lens. To distinguish between the $d_i$ on the left and the right side of the lens, we will express the virtual $d_i$ (on the left side of the lens) as having a negative value where necessary. Note that this difference is not crucial to the general line of reasoning. Another point to stress is that we included $f$ in above reasoning. Obviously, $f$ is not directly observable as an

---

6. As there is not much sense in trying to explore the effects of $d_i$ when $\alpha = 0$. Therefore, we will continue assuming that $\alpha \neq 0$.

7. In this chapter, "proportionality" (forbus, 1984) implies qualitative proportionality, or, in mathematical terms monotonicity.
important point and its importance has therefore to be induced from the generalities of behavior in the domain. As without the inclusion of this concept all the following derivations cannot be made, including this point in the reasoning process is crucial to successful discovery.

Based on the above described observations, the following five patterns can be inferred:

2. if \((d_o > f)\) then \(d_o\) is inversely proportional to \(d_i\)
3. if \((d_o < f)\) then \(d_o\) is proportional to \(|d_i|\)
4. if \((d_o\approx f)\) then \(d_i\approx \infty\)
5. if \((d_o < f)\) then \(d_i < 0\)
6. if \(d_o\approx 0\) then \(|d_i|\approx 0\).

As in these five regularities \(d_i\) is always a related to \(d_o\) but modulated by the relation between \(d_o\) and \(f\), a promising start for a formula describing these regularities takes the form \(d_i = f(d_o, f)\). Given the above presented patterns, one can infer that:

- Based on Pattern 2, \(d_i\) might be based on a division by \(d_o\), as if \(d_o\) gets larger, \(d_i\) gets smaller and vice versa.
- Pattern 3 suggests that \(f\) functions as a threshold, if over this threshold, the effect inverses.
- Pattern 4 also hints at a division, as dividing by zero yields infinity. Given the conditions of Pattern 4 and Pattern 3, and the assumption that \(d_o\) is part of the denominator, both \(d_o - f\) and \(f - d_o\) are candidates for the denominator part of the formula defining \(d_i\).
- Based on Pattern 5, the first of these two is correct as if \(d_o\) becomes smaller than \(f\), the resulting \(d_i\) drops below zero.
- On the basis of Pattern 6, the numerator can also be more accurately described. As a division only results in zero as the numerator is zero, the numerator should contain a expression that evaluates to zero if \(|d_o|\) approaches zero (e.g., \(d_o \times \ldots, \sin(d_o)\), etc).

Based on these derivations, the formula describing \(d_i\) can be specified as:

\[
d_i = \frac{g(d_o, \ldots)}{(d_o - f_i)}
\]

To decide on the nature of the function \(g(d_o, \ldots)\), more information is necessary. If
quantitative information is available, \( g(d_o, \ldots) \) shows to be a simple linear function dependent on both \( d_o \) and \( f \) (c.f., Van Someren & Tabbers, 1998). However, even without quantitative data the correct derivations can be made.

Given the observation that when \( d_o \) is increased, \( d_i \) decreases but at a rate negatively correlated with the size of \( d_o \). This would suggest asymptotic behavior for large values of \( d_o \). This can be reformulated as in Pattern 7 below:

7. If \( d_o \) approaches \( \infty \), \( d_i \) approaches \( n \)

Given the above presented formula for \( d_i \), if \( d_o \) becomes infinite, the denominator would also evaluate to infinity as \( f \) effectively drops out of the equation. Therefore, to arrive at \( d_i = n \), the numerator has to be \( n \times \infty \), that is, \( n \times d_o \) as it was already known that \( d_o \) has to be part of the numerator and is infinite. Given the observation that both \( n \) and \( f \) differ per lens, and appear to be highly correlated, expressing the numerator as \( f \times d_o \) is worth a try.

This lead to the following formula:

\[
d_i = \frac{(d_o \times f)}{(d_o - f)}
\]

which can be rewritten into the more commonly known lenses law:

\[
\frac{1}{d_i} = \frac{1}{f} - \frac{1}{d_o}
\]

Obviously, although based on observations done in the simulation, one cannot prove the correctness of this law without resorting to quantitative measuring. For example, the last step (treating \( n \) to be equivalent to \( f \)) can only be validated based on quantitative data.

However, the above reasoning shows that even in the relative simple configuration presented in the first phase of Prins’ Chapter 5 experiments, the lenses law can be derived without having to resort to quantitative measures. However, it was necessary to incorporate information that was gained by observing the same qualitative state in different quantitative configurations. Nevertheless, this shows that the knowledge that can be derived from qualitative reasoning combined with within state modifications suffices for the main discoveries in the Optics domains. To distinguish between a
pure qualitative model and a qualitative model in which quantitative modification can play a role, I will use the term conceptual model to refer to the latter.

**Behavior Analysis**

**Analyzing Post-test Behavior**

**Applying the Conceptual Model** Now that we have shown that the conceptual model in itself is sufficient to discover the knowledge embedded in the Optics simulation, we can apply this conceptual model to the learners' behavior in the Optics simulation. That is, for each (dynamic) experiment-state change, we can infer what knowledge might be derived by the learner from that change. This way, we can conduct an analysis analogous to the one conducted for the Peter-task: focusing on the consistency between the learners' actions during the discovery sessions and the scores on the post-test. As concluded for behavior shown in the Peter-tasks, we would expect this analysis to show that having seen a certain state-change is correlated with the answers at the post-test.

Ideally, this consistency analysis should be done automatically, as the simulation environment "knows" in which state the configuration is and also could know the answers to the questions the post-test. However, as this is a post-hoc analysis, the scoring of learner's behavior had to be based on the recorded action-logs. Although scoring could have been based on the 74 dynamic experiment-state changes, some of these states are less informative with regard to determining the quality of the discovery learning process. Therefore, the focus was on a subset of dynamic experiment-state changes discussed in next paragraph.

**The Scoring Scheme** The lighter colored information in Figure 6.10 is a schematic representation of a lenses configuration as shown in Figure 6.3. Each black arrow represents an action that is scored. For example, the top-most vertical arrow is associated with a change in the angle of the light beam causing the light beam to move out of the lens. Each of the vertical arrows represents a change in the angle of the light beam. The horizontal arrows represent both movements of the lamp and of the lens. As most learners had a tendency to either move the lamp or move the lens, learners seem to infer or know that moving the lamp has the same effect as moving
As discussed earlier, Optics is a continuous domain. Changing the angle of the light beam does not necessarily result in a single observation as learners might have observed intermediate situations. Algorithms in the Optics module responsible for outputting the log file information discretized these movements. If the movement was interrupted or slowed down, this was reflected in the Optics log file by a separate entry; quick movements are not logged. This mechanism allows objective measurement of the continuous manipulations of the learners; each event written to the log file is taken as data-point for analysis. Appendix 5.A shows the scores per learner per type of action scored.

**Scoring and Interpreting the Post-test** In the Peter task the constructed experiments are relatively salient; as learners have invested effort to construct a particular experiment, the rationale behind the analysis in Chapter 3 and 4 assumes that learners have perceived the constructed experiments. However, this assumption might not hold in the Optics task. First, a particular configuration might be nothing more than a “transfer configuration” from one perceived state to another. For example, if a
learner currently observes a light beam hitting the lens above the optical axis and wants to construct an experiment in which a light beam intersects with the lens below the optical axis, the learner will encounter a situation in which the light beam goes through the heart of the lens. In other words, one can not be sure that all constructed configurations have actually been perceived separately like the experiments in the Peter-task. Second, even if a learner observes a certain situation, we cannot assume what knowledge would be derived from that observation as the learner might not be aware of the important variables or quantity spaces to pay attention to.

Still, the assumption that the post-test measures the quality of the discovery process implies that the learner should have perceived the configurations that could be at the basis of knowledge derivation. Applying the consistency-based analyses as presented in Chapters 3 and 4 to the current setting, one would expect that the relative low scores on the post-test are at least partly explained by just not having observed the appropriate configurations.

In the Peter-task, the post-test scores were compared to what could have been derived from the constructed experiments assuming that the learner had used optimal reasoning skills. Although the mapping from observed configurations to knowledge is less clearcut in the Optics task, the conceptual model provides a framework that enables a similar analysis. Instead of conducting a bottom-up analysis, basing our analysis on the conducted experiments, we conduct a top-down analysis in which the post-test questions determine the configurations to focus on.

For each of the questions in Prins’ experiment, a set of dynamic experiment-space states was identified that in the GARP model led to the behavior that was queried for. This way, the consistency can be assessed by calculating the correlation between the answers to the post-test and the number of observations made in the associated set of “critical states”.

However, an additional problem with the Optics domain is that all the learners in the current experiments have had some prior but limited formal Optics schooling. In the Prins’ experiment, this schooling was at least four years prior to enrolling in the experiment. Therefore, most learners were at best aware of the task; none could be considered expert. Still, learners might remember particular effects from this prior

\(^8\)Note that this approach was possible because the post-test for the Peter-task queried the learners for all possibly derived knowledge.
formal education. As all the post-test questions had also been presented during the pre-test, the results on the pre-test can be used to select to which questions learners might have known the answers. Obviously, learners might have correctly guessed the answers to the pre-test questions. Therefore, one might either argue for including all questions in the analysis or excluding those that have been answered incorrectly during the pre-test. If one would include all questions in the post-test, regardless of the associated answer on the pre-test, this would lower the validity of the test. That is, learners might decide not to construct experiments for a certain phenomenon that they already know about. Therefore, the lack of observation associated with a correct answer will lower the correlation. If only the questions that are answered incorrectly on the pre-test are included, some learners' data will not be included in the analysis as they accidentally might have guessed the correct answer. However, this will not influence the statistic, but merely lower its power.

Based on the above considerations, the 32 analyzed learners (selected out of a total of 44 learners, see Appendix 5.A) combined with 15 post-test questions yielded a total of 287 observation counts versus post-test score pairs for analysis. An issue with using the raw number of observations is that it does not distinguish between a very systematic experimenter who does a small number of carefully constructed experiments in a certain experiment space and a more sloppy experimenter who created a large number of experiments, and therefore multiple experiments in that same experiment space. When using the raw number of observations, the sloppy experimenter's count diminishes the effect of the systematic experimenter. Therefore, all raw number-of-observations have been rescaled to a zero for either zero or one observations or a one for two or more observations.

The overall correlation between the number of observations in the critical states and the related question in the post-test is 0.18. Although this statistic is low if compared to other correlation based statistics, a correlation of .20 is associated with a 60% overlap in scores in this type of test of two binary variables. To test the general significance of this effect, a t-test was conducted over the correlations per question between the binary number-of-observations and the binary post-test scores. The results showed that this correlation was indeed significant (t=2.10, df=14, p<0.05). However, this effect is slightly distorted by a number of questions which had a very high number of correct pre-test answers. Therefore, only a small number of learners
contributed to the correlation for that particular post-test question. When values that are entered in the t-test are weighted for the contributing number of learners, the effect becomes more pronounced (t=8.83, df=14, p < 0.00).

This indicates that having observed a transition predicts the associated answers on the post-test – even in a relative opaque task as Optics in which the important variables are not as clear cut as in the Peter-task,

Nevertheless, question remains why this correlation is so much lower then the correlations observed in Chapter 3 and 4. Table 6.3 shows a contingency table that is based on the same data as above correlations. Applying a $\chi^2$ analysis showed that there is indeed an effect of having seen post-question related configurations ($\chi^2 = 8.38, df = 1, p < 0.01$). However, the table also clearly shows why the above correlation is relatively low. For a sizable number (76) of incorrect post-test answers, learners had seen the relevant configurations but answered the associated post-test question incorrectly. As argued before, this might be due to opaqueness of the Optics task: Even if one observes that the angle of deviation is inverted when comparing two configurations with lamps at different x-positions, one still has to infer that there is a special point where the refraction qualitatively changes.

<table>
<thead>
<tr>
<th>No. of relevant observations per post-test question</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\leq 1$</td>
<td>$&gt; 1$</td>
</tr>
<tr>
<td>Answer incorrect</td>
<td>80</td>
<td>76</td>
</tr>
<tr>
<td>Answer correct</td>
<td>44</td>
<td>87</td>
</tr>
</tbody>
</table>

Table 6.3: Overview of the number of times a post-test question was answered correct or incorrect separated for whether or not the learner had seen more than one configurations related to that question.

ANALYSIS OF THINK-ALOUD PROTOCOLS

Although the mapping between behavior and scores is not as clear cut as in the Peter-task, the analyses of the Optics task show that there is a relation between what the learners have seen and to which questions they know the answers. Although this result is straightforward, it does show that answering questions incorrectly is not necessarily caused by suboptimal discovery skills, but shows that is might as well be caused by simply not having encountered the situations from which post-test related knowledge could be derived. If the low scores are partly due to not having seen
certain important states, the regularities reported by the learners should be valid for their observed states. In other words, it might be that they inferred knowledge that is only locally correct: Knowledge that might generalize of the observed experiment, but that does not hold for all possible states. To assess the validity of this hypothesis, we will discuss the think aloud protocol of two of the learners in more detail.

Learner reads the instruction, places lamp and double convex lens on workbench:

Nothing happens. Or it might go straight, that would also be a possibility. Let’s reposition the lamp. Hmmm. Rotate the lamp, euhm, light beam. When you rotate it, then it doesn’t go straight anymore, then it goes down. The thick lens refracts downwards. OK. Remove the lens. Flat-convex lens, light on, that one refracts as well, refracts less strong. Remove. Double-concave lens. Light on, and that one refracts upwards. Hmmm. A surprise... Also removed. Flat-concave lens, that one will probably refract less. Euhm, and that seems to be correct, let’s test that once more. Yes, that is correct. Ok, that is everything I need to know.

Learner rereads the instruction, realizes that the part in the assignment about “what determines the location of the intersection between optical axis and light beam” has not been addressed yet:

Euhm, the light beam does not intersect with the main axis when the lens is a negative lens, I’ve already discovered that. [...] And when a positive lens is used, the location is determined by whether or the lens is flat-convex or double-convex. When it is double, the intersection is closer by, and probably also by moving the lamp. OK. Double-convex lens. Light on. Lamp closer by, euhm, lamp closer by means that it intersects further. OK.

The knowledge that this learner has gathered is limited to knowing that lenses with one flat side refract less, that concave lenses refract in “the other direction” and that by “moving the lamp, the intersection is closer”. Unsurprisingly, this learner has difficulties answering some of the post-test questions.

9The instruction given to the learners in the experiment under study (Prins, 2002, Chapter 5) was relatively detailed compared to the other Optics studies. The learners were instructed to discover what determines the location of the intersection between the refracted light beam and the optical axis.
**Discovery skills in a complex task**

Learner reads the questions out loud and then answers the question:

The light beam is pointed down a little, where shall the light beam hit the optical axis? Hmmm. So that's something that I didn't look previously. Euhm, it has to be based on something that I remember from school. Hmmm, closer to the lens.

However, although this learner is not able to answer all the post-test question correct (especially because the recalled secondary school knowledge is incorrect), the learner is able to induce regularities from observed phenomena and seems to posses accurate discovery skills, even in the complex Optics environment. For example, after encountering a unexpected result, the learner is able to overcome the initial surprise and incorporates the new finding. Even more, the learner states hypotheses, induces regularities and even tests a discovered effect in a different configuration to test its generality. One could state that the discovery behavior of this learner is per definition below par as only a limited set of observations is generated. However, as the think-aloud protocol of the next learner shows, similar behavior is also observed in learners of which the behavior is more extensive.

Learner reads the instruction, places lamp and double convex lens on workbench:

Now I should look what happens through that lens. Well, when I let it go straight through, then it goes straight, now I'll move the lamp, it should not matter whether I move the lamp because it just keeps going through the heart of the lens. [...] OK, but I can rotate the light beam, so, rotate upwards [...] when I make the ingoing angle larger, the outgoing angle also becomes larger, let's see if that is correct, no, of course not. This angle is much larger. With such a double-convex lens, the smaller the angle of incidence, the larger the outgoing beam, what happens when I move in the horizontal plane, yep, that angle remains of course the same. Yes, the further the lamp from the lens, the [...] smaller the outgoing angle. That is, the smaller, when it becomes smaller, this angle will become larger, and that is not true as this is a very small angle. So... Euhm. The smaller the ingoing beam, the smaller the outgoing beam.

The learner changes the double-convex lens for a flat-convex lens, conducts a series of experiments, and confirms above conclusions.
Now I'll start looking what will happen when I reposition the lamp and lens relative to each other. If I move the lamp closer, then this angle becomes, hmmmm, the closer by the lamp, the smaller that angle. [...] It just occurred to me that one can also pay attention to how far the outgoing light beam is from the lens. So, I'll check... the closer the lamp to the lens, the further the outgoing beam hits the main axis. [...]  

The learner removes all attributes from the workbench, does some experiments with negative lenses ("The refraction is, compared to my intuition, let's say, in the wrong direction") and then starts to reposition the lamp:  

Now I'll put the lamp closer by and further away from the lens, so of course the angle will remain the same. That is, the closer, euhm [...] euhm, the further the lamp, the larger the angle, does that also hold for the other, the further the lamp, the – oh, no, I've been doing the closer, euhm [...] how can I say this [...] with the last we've seen that the beam intersects with the main axis, but this beam doesn't do that, so I'll have to rephrase something so that I can remember it more easily... the closer the lamp to the lens, the straighter the beam is”  

Learner notices that 20 minutes have passed, and makes some remarks that it will be difficult to finish all phases within 1.5 hours.  

Next lens, [...] I'll bet that it will show the same effect [...], and yes, the same thing happens, for the double-convex lens, it was the larger the incoming angle, the larger the outgoing angle, changing the horizontal position, I'll move the lamp closer to the lens, the closer the lamp to the lens, the more straight the beam goes. Yes, that's right. Euhm. Well, I think it will be same for the other lenses, let's continue to the next phase.  

The outcome of this discovery behavior is clearly suboptimal, as – again – the resulting knowledge is only correct for a subset of the total experiment states. However, this learner – even more than the learner discussed above – did show to possess the capabilities of experimentation and of deriving knowledge from experiments. Moreover, the learner also shows (based on the manipulation of the light beam angle) that she knows about the “search for extremes” heuristic, which is considered to be essential in describing a complete domain. However, either by bad luck or by a not
vocalized decision, this learner only applies this search for extremes while searching for an effect of the angle/position of incidence, not for the location of the lamp. Nevertheless, the learner shows to have all prerequisites to become a successful discovery learner. That she used the search for extremes heuristic in a suboptimal part of the experiment space is indicative of a lack of knowledge, not of a lack of skill or impoverished behavior. Therefore, if the learner had made the post-test with the reported knowledge, the resulting meager post-test score would have been misleading for the quality of this learner’s discovery skills.

The type of analysis used in this chapter can be seen as a combination of the ideas of Chapter 3 and 4 and the sequence based analysis method used by Hulshof (2001). In his thesis, Hulshof identified low-level sequences of behavior and correlated these with post-test outcomes. The method used in this chapter also looks at behavior in small sequences - but is knowledge-richer, and therefore allows researchers to focus on specific sequences of behavior within states.

CONCLUSIONS

Limited effects of discovery learning are partly due to not having explored the regions where important discoveries can be made. Apart from this conclusion, which is in line with the one presented in Chapter 3 and 4, another conclusion that can be drawn from above presented data is related to how learners infer regularities.

Without inferring the regularities, learning would be nothing more than remembering instances, which would imply extensive demands on memory. Moreover, the instruction given to the learners at the beginning of each phase also implies that there are single regularities to be learned (e.g., “What determines the location of the intersection between outgoing light beam and optical axis?”). However, learners have difficulties finding the “right” hypotheses. If we assume that to create good hypotheses learners have to collect observations which have different values for the main variables, the first difficulty the learners face is to identify these variables. Only after having identified these variables, learners can start to collect the right set of observations. Because of this, discovery in complex domains like Optics is a two-step process. In contrast, the Peter-task does not require this first step to take place. The interface and instruction of the Peter-task already identify the variables of interest. Based on this information, constructing VOTAT experiments and inducing knowledge from
those experiments immediately leads to the right level of knowledge.

As was shown earlier in this chapter, it is possible to derive the lenses law from purely qualitative knowledge. However, this derivation was facilitated by knowing what the outcome of the derivational process had to be, and it was therefore possible to ignore irrelevant outcomes. Learners do not have this advantage, and if they are not given explicit information about what to look for, all behavior need to be scrutinized to assess whether that particular type of behavior might be key to discovering an important regularity.

This notion of discovery as a two-step process can also be part of the explanation of the difference in performance between the Hulshof and the later Prins studies as presented in Table 6.1. Because of the more focussed instruction and the more constrained experimental settings in the Prins studies, identification of important variables becomes feasible. In contrast, the instruction used in the Hulshof’s Optics settings was explicitly stated in very broad terms, not focusing the learner to certain variables or behavior. The sole purpose of the instruction in these studies was to get the learners working in the environment (Hulshof, personal communication). Although this instruction probably leads to the least guided (and therefore arguably the most natural) type of discovery learning, the gain of discovery was low in these studies.

Given that these differences yield a large performance effect, combined with the finding that learners are able to deduce the lenses-law when they are given an appropriate dataset (Van Someren & Tabbers, 1998), supports the conclusion that the domain was not too difficult. Instead, the meager results might be associated with the task. To assess this hypothesis, we compare these tasks on four dimensions: experimentation, variable identification, conceptualization and complexity.

One of the differences between the Optics studies was the relative difficulty to generate appropriate experiments. In the Van Someren and Tabbers study, the learners where presented an appropriate set of experiments. In the later Prins studies, the instruction and task-setup provided support in constructing the right kind of experiments. In the Hulshof studies, the broad instruction provides learners with limited guidance what experiments might be useful. This makes it even more likely in the Hulshof studies that learners have conducted experiments in suboptimal experiment spaces, as learners showed great experimentation variability even in the more guided
Prins studies.

But even in the Prins' studies, to come up with the kind of knowledge assessed at the post-test, learners had to identify the important **variables** and their quantity spaces. Only after discovering the quantity spaces and realizing their importance, learners are able to construct the right kind of experiments. Given the relative ease with which the learners in the Van Someren and Tabbers study inferred the lenses-law, this factor seems of utmost importance in explaining the difference in performance between that study and the ones by Hulshof and Prins.

A third factor that distinguishes the Van Someren and Tabbers study and the Hulshof and Prins studies is the amount of **conceptualization** necessary. The Van Someren and Tabbers study provided the learners with an appropriate background story and a template for explaining the presented results. Only a relative small part of the task was actual conceptualization; in a sense it can be seen as a formula deduction process. In contrast with that study, learners in the Optics simulation had to decide what the appropriate type of knowledge was that had to be derived from experimenting. For example, is a simple regularity sufficient, or does one need to infer a mathematical relation? And, related to that, what kind of relation to expect? In the Peter-task presented in Chapter 3 and 4, the majority of learners did not discover the second-order effect. Similarly, most of the learners in the Optics task did not come up with the right kind of relations describing the different variables. For example, the continuous character of experimentation might have induced a conceptualization of continuous behavior between the important variables. However, given that most of the learners did not incorporate the virtual focal point in their reasoning, continuous behavior is an incorrect assumption. This way, the task setup actually supported an incorrect conceptualization of the domain.

The fourth factor is **complexity**. Obviously, the three factors discussed above all contribute to the classical notion of complexity. For example, the harder it is to find the proper conceptualization, the more complex a task is. However, another notion of complexity is whether it is possible for a learner to **think** that a set of regularities has been found that covers the whole experiment space. Because of the given data set in the Van Someren and Tabbers study, this was not an issue in their study. In the Hulshof and Prins studies, however, learners reported to be ready experimenting in a certain phase, only to discover during the post-test questions that their knowledge
did not cover the complete system.

Given the above reasoning, one can argue that learners actually perform remarkably well in the complex Optics simulation. Although they have to construct interpretable experiments, which is a feat in itself, although they are initially not aware of some of the important variables, although they do not know the correct conceptualizations and are not aware of all the complexities involved, learners are able to counter all these issues and induce regularities. And despite the incompleteness of the found regularities, most often they are correct for the subset of behavior observed by the learner. Therefore, regardless of the scores on the post-tests, learners in the Optics task do show remarkable discovery skills. Expecting learners to do better in this task, that is, to derive regularities at the level of the lenses law, might be best described as an underestimation of the challenges learners face during Optics experimentation.
APPENDIX 5.A

The table below shows, per learner, the number of manipulations for each of the state changes shown in Figure 6.10. As each of the four different lenses has its own set of state changes, the second column of this table denotes the lens during which the state changes were made. (DCV maps to double convex, DCC to double concave, SCV to single convex and SCC to single concave lens)

<table>
<thead>
<tr>
<th>Learner</th>
<th>Lens</th>
<th>Position Lamp</th>
<th>Intersection Beam/Lens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LoVF P</td>
<td>RoVF P</td>
</tr>
<tr>
<td>201</td>
<td>DCV</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>202</td>
<td>DCV</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>203</td>
<td>DCV</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>204</td>
<td>DCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>205</td>
<td>DCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>206</td>
<td>DCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>207</td>
<td>DCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>208</td>
<td>DCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learner</td>
<td>Lens</td>
<td>Position Lamp LoVFP</td>
<td>RoVFP</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
<td>---------------------</td>
<td>-------</td>
</tr>
<tr>
<td>209</td>
<td>DCV</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>210</td>
<td>DCV</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>211</td>
<td>DCV</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>13</td>
<td>.</td>
</tr>
<tr>
<td>212</td>
<td>DCV</td>
<td>4</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td>3</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>11</td>
<td>.</td>
</tr>
<tr>
<td>213</td>
<td>DCV</td>
<td>6</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td>2</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td>214</td>
<td>DCV</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td>6</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>215</td>
<td>DCV</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>216</td>
<td>DCV</td>
<td>2</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>3</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td>217</td>
<td>DCV</td>
<td>.</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>.</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>.</td>
<td>9</td>
</tr>
<tr>
<td>218</td>
<td>DCV</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td>219</td>
<td>DCV</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
### Discovery skills in a complex task

<table>
<thead>
<tr>
<th>Learner</th>
<th>Lens</th>
<th>Position Lamp</th>
<th>Intersection Beam/Lens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LoVFP &lt; -&gt;</td>
<td>RoVFP &lt; -&gt;</td>
</tr>
<tr>
<td>220</td>
<td>DCV</td>
<td>. .</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>. .</td>
<td>. .</td>
</tr>
<tr>
<td>221</td>
<td>DCV</td>
<td>2</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>DCV</td>
<td>3</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>2</td>
<td>. .</td>
</tr>
<tr>
<td>223</td>
<td>DCV</td>
<td>41</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>DCC</td>
<td>7</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>4</td>
<td>. .</td>
</tr>
<tr>
<td>224</td>
<td>DCV</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>DCC</td>
<td>5</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>5</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>.</td>
<td>. .</td>
</tr>
<tr>
<td>225</td>
<td>DCV</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>DCC</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>226</td>
<td>DCV</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>DCC</td>
<td>.</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>1</td>
<td>. .</td>
</tr>
<tr>
<td>227</td>
<td>DCV</td>
<td>.</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>DCC</td>
<td>.</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>.</td>
<td>. 1</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>.</td>
<td>. 1</td>
</tr>
<tr>
<td>228</td>
<td>DCV</td>
<td>.</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>DCC</td>
<td>.</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>.</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>.</td>
<td>. .</td>
</tr>
<tr>
<td>229</td>
<td>DCV</td>
<td>.</td>
<td>. .</td>
</tr>
<tr>
<td>230</td>
<td>DCV</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>DCC</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>231</td>
<td>DCV</td>
<td>.</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>DCC</td>
<td>1</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>.</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>.</td>
<td>. .</td>
</tr>
<tr>
<td>232</td>
<td>DCV</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>DCC</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>SCV</td>
<td>4</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>6</td>
<td>11</td>
</tr>
</tbody>
</table>
CHAPTER 6.

COMPONENTS OF INDUCTIVE LEARNING

Abstract

In this chapter, four different dimensions are presented along which inductive learning tasks can be compared. Also, we will discuss to what extend the conclusions from previous chapters of this thesis are applicable to inductive learning in general. Based on this discussion, three factors are presented that explain both successful and less successful results on inductive learning tasks. This chapter concludes with suggestions for further research and with a discussion of the implications of this research for the application of inductive learning in the curriculum of modern education.

In the introduction of this thesis, inductive learning was compared to scientific discovery, as is common practice (e.g., Klahr & Dunbar, 1988; Klahr, 2000; Kuhn et al., 1995; Schunn & Anderson, 1999). The assumed correspondence between scientific discovery and inductive learning implies a central role for hypothesis testing in inductive learning (c.f., SDDS, Klahr & Dunbar, 1988). Given this centrality of hypothesis testing, conducting experiments is not so much a task in itself, but a necessity for constructing, testing and refining hypotheses. The studies presented in this thesis, however, shed a different light on inductive learning.

In contrast to the notion of a central hypothesis, inductive learning in the tasks presented in this thesis appears to be mainly focused on the construction of experiments
to “test the effect of that variable”, without initially having a carefully constructed hypothesis. When learners construct hypotheses, these are often based on a simple evaluation of their prior knowledge. Instead of criticizing this behavior as imperfect or (too) shallow, this concluding chapter, in line with the reasoning in previous chapters, presents a rationale for this behavior and explains it in terms of bounded rationality (Simon, 1957).

To generalize the conclusions of this thesis to the domain of inductive learning, it is necessary that the conclusions are based on a set of inductive learning tasks instead of being specific to one particular task. The set should reflect a range of inductive learning tasks that is as broad as possible. In the next section of this chapter, we present four different dimensions along which inductive learning tasks can be compared and argue that the tasks discussed in this thesis cover a broad range of the spectrum of inductive learning tasks. Then, we will discuss the tasks presented in this thesis focusing on the conclusions from previous chapters that are suitable for generalization to inductive learning in general. Based on these conclusions, three factors are presented that explain both successful and less successful results on inductive learning tasks. Finally, we will conclude with suggestions for further research and with a discussion of the implications of this research for the application of inductive learning in the curriculum of modern education.

**FOUR DIMENSIONS OF LEARNING TASKS**

Chapters 2 to 5 discussed inductive learning in different domains using different tasks. Dimensions of inductive learning tasks that determine the difficulty of the task include the extent to which learners need to perform experimentation, variable identification and conceptualization, and complexity. The tasks are compared on these dimensions to ground the assertion that they form an appropriate subset of the complete spectrum of inductive learning tasks.

**Experimentation** The tasks differ in the availability of the data on which conclusions have to be based. In the balance scale task, the problems presented to the learners are selected by the experimenter, the learners cannot construct experiments by themselves. This constrains how easily a learner can test constructed hypotheses, as the presented experiments might not be the experiments the learner needs to test a current hypothesis. On the other hand, this could also be
Components of Inductive Learning

an advantage, as a carefully constructed sequence of experiments might guide the learner in the right direction for deriving new knowledge (c.f., Chapter 2). In both the Peter-task and the Optics tasks, learners have to construct experiments. Although this enables learners to actively test their hypotheses, it also burdens them with the task to construct correct experiments and the decision about when to stop creating experiments.

Variable Identification The variables that underly the simulation behavior in the tasks can be either relatively easy identified from the task or need to be actively induced by the learner. With respect to the identification of the variables, it is not only important that the learner identifies a particular representation in a simulation as an important variable, but also identifies its structure, i.e., its quantity space (see Chapter 5). In the Peter-task, both the variables and their quantity spaces are easily identified. Both the five variables and the quantity spaces of these variables are given by the interface. In the balance scale task, only two aspects of the balance scale are modified over different representations, indicating that these are useful variables for the inductive learning process. However, these variables are not identified as significant variables in either the instruction or the task-setup. Therefore, part of the difficulty of this task is to identify that these variables play a role in determining the behavior of the system. If these two variables are selected as interesting, the interface prescribes their quantity-spaces (i.e., the maximum number of weights that can be placed on a single peg, and the number of pegs for the maximum distance). As discussed in Chapter 2, the identification of variables is an important aspect in the explanation of children's behavior on the balance scale task. However, for the third task, Optics, both aspects are not straightforwardly derived. Although all potential variables are visible in the interface, the learner has to select the correct variables for his or her experiments and derivations and has to induce the quantity space to derive correct conclusions.

Conceptualization Although the identification of the relevant variables is an important, independent part of the inductive learning process, it is also a prerequisite for a correct conceptualization of the task. Conceptualization refers to finding the relevant set of concepts to describe the behavior of the domain under study, for example, the type of relations between the identified variables, but
also more generally how the learners think about the task.

In the Peter-task (and especially in the Peter-goes-shopping variant), the task-format and the associated instruction direct the learner toward an adequate conceptualization. Learners have no problem identifying that they have to discover the effect of the different levels of the shown variables on a pre-specified outcome. Even if the existence of an interaction effect is not known to the learners, the task-setup or domain appears to guide the learner toward a correct conceptualization of interactions (i.e., a conditional-effect).

However, in both the balance scale task and in Optics, the learner has not only to identify which variables are important, but also has to identify the type of relation that links different values of an independent variable to an outcome on a dependent variable. In the balance scale task, the learner has to come up with the idea to multiply the values for weights and distances (even if the often earlier thought of addition of weights and distances also proves relatively successful). In Optics, the learner has to induce that most of the effects are, qualitatively speaking, discontinuous. (That is, both the virtual focal point and the heart of the lens are points in quantity spaces in which the behavior in the domain shows a discontinuity.)

**Complexity** As in all domains, different inductive learning tasks have different levels of complexity. However, it is difficult to judge what defines complexity of a task as there are multiple, not necessarily related, aspects that define it. Here we will not focus on task aspects in itself, but on how these task aspects are perceived by the learners.

In all three tasks presented here, a learner might come up with a set of relations that appears to give a correct or at least an adequate description of the domain, whereas in fact this description is only valid for a particular subset of the domain. Although a description that covers the complete domain is obviously the best description, the partial description might be correct for the subset of data the learner has seen. Take, for example, the balance scale task. If a learner uses the Addition Rule and does not encounter any items for which the Addition Rule renders the incorrect answer (e.g., the item with two weights on peg two on the one side, and four weights on peg one on the other side), the used Rule fits the data perfectly. Similar examples can be found in the two other
Components of Inductive Learning

tasks, for example, the main effects without interactions in the Peter-task and experimenting without moving the lamp over the virtual focal point in Optics.

Obviously, one can find such a subset in almost all relatively complex tasks. However, complex tasks can differ in the ease with which learners are able to infer whether their current description fits the complete domain. An important concept in this respect is whether or not a task is ill-defined. All three tasks presented in this thesis are ill-defined in that it is not clear what the outcomes of the discovery process should be, nor which variables should be incorporated in the discovery process. Therefore, learners need to use heuristics. For example, because of the ill-definedness of the variables, learners do not know what the complete experiment-space is. During experimenting, they need to assess whether they have covered a sufficient area of this unknown experiment-space to warrant generalizing their conclusions. But not only does this influence the experiment-space, the ill-definedness also influences the hypothesis-space as learners do not know what type of relations need to be searched for.

Because of the differences in ill-definedness per task, the tasks also differ with respect to how salient the incompleteness (compared to a complete description) of a certain set of observed regularities is. For example, although the Peter-task has a more complex set of underlying regularities, these regularities have a simpler structure than the multiplication rule in the balance scale task. Related to this is the relative gain of constructing experiments; in some tasks it is easier to construct experiments that refute the current set of regularities than in others. For example, it is relatively easy to reject hypotheses in BigTrak (Klahr & Dunbar, 1988), as almost all incorrect hypotheses have a limited scope of accurate predictions. On the other hand, if a learner does not know about the virtual focus point, a lot of experiments can be constructed that support an incorrect hypothesis.

As the balance scale task is relatively well defined; it is known to the learners what to predict and the variables of interest are easily identified (by adult learners), the main difficulty is finding the right method for combining the weights and distances. The Peter-tasks are one step more complex, in that in these tasks the learners have to detect which variables are actually influencing the outcome. Moreover, the right description can only be found when a learner constructs
the right set of experiments; there is nothing in the task that alerts the learner to this set. The most complex of the three tasks discussed is the the Optics task. In this task, the variables are identified, but not the type of knowledge that learners are expected to gain nor the quantity spaces (see Chapter 5) of the variables. And, similar to the Peter-task, learners can come up with a partial description that accurately fits a large set of observed experiments while not being hinted by the conducted experiments that the current description is inaccurate.

Table 7.1: Categorization of the inductive learning tasks presented in this thesis.

<table>
<thead>
<tr>
<th>Experiments</th>
<th>Balance scale task</th>
<th>Peter task</th>
<th>Optics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-directed</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Externally generated</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variables (QS)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-induced</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Visible in interface</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Conceptualization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-induced</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provided by task setup</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Complexity</td>
<td>low</td>
<td>intermediate</td>
<td>high</td>
</tr>
</tbody>
</table>

Table 7.1 gives an overview of the position of the tasks on the four dimensions.

THREE TASKS

Even though the three tasks discussed in this thesis cover the complete spectrum of the above presented dimensions, the inductive learning behavior in these tasks shows a relatively large overlap.

BALANCE SCALE TASK

In the chapter on the balance scale task, a computational model is presented that explains how children's behavior on this task develops from using a simple guessing rule to using the correct rule which involves calculating the torque. A central feature of the explanation of development is prior knowledge about solving forced choice tasks. If the model is presented with a forced choice task, it has access to knowledge
Components of Inductive Learning

that states to search for a difference between the alternatives of the forced choice. In the context of this task: Search for differences between the right and left arm of the balance scale. If the model finds a difference, this difference is used in solving the problems and new knowledge is constructed that relates those problems to that difference. Initially the model uses only one type of difference (i.e., the difference in the number weights between the left and right side of the balance scale). This causes a relatively large number of erroneous predictions about the movement of the balance scale. However, as searching for differences is resource intensive, the model does not immediately search for a new difference if an erroneous prediction is made. Only if the number of erroneous predictions becomes too large compared to the effort that it takes to find a new type of difference, the model will invest in a search for these new differences. This architecturally implemented mechanism leads to a model that performs according to satisficing principles: Do not search for new explanations as long as the current knowledge is not too often falsified.

PETER-TASK

In Chapter 3, the first chapter on the Peter-task, computational models are presented that focus on conducting experiments instead of focusing on constructing complex hypotheses as is put forward by theories like SDDS (e.g., Klahr & Dunbar, 1988). Nevertheless, without the emphasis on hypothesis construction or testing, these models do capture the main behavioral patterns. This illustrates that instead of a complex search for the correct type of hypothesis, inductive learning performance in a simpler task like the Peter-task is constrained by the learner's ability of constructing correct experiments and deriving knowledge from these experiments. But what guides the construction of the experiments? In the Peter-task, the main determinant is the computer-interface. Most learners proceed by starting with the top-most variable working their way downward. However, another determinant is prior knowledge. Learners also construct experiments based on an evaluation of their prior knowledge. If they assume a particular effect to be associated with a level of a variable, they will explicitly test this level. Moreover, if an effect is discovered that is not in line with their prior knowledge, this is often reason for a more thorough investigation of the effects underlying that particular level or variable. In the strongest form, this leads to behavior that limits inductive learning to what is covered by the experiments that are “dictated” by the interface, supplemented with experi-
ments constructed on the basis of prior knowledge. If the behavior satisfied these two constraints, learners often see no reason to continue learning although aware of the not conducted experiments. This can be seen as an illustration of the satisficing principle: Although there is maybe more to be known about the task, if there are no more loose ends, the inductive learning process is considered finished.

Chapter 4 takes this one step further and proposes a measure of consistency as an alternative measure for the quality of the inductive learning process. Based on a comparison of two Peter-experiments, it was shown that the score on the new consistency measure is more stable over domains than the previously used measures. These more stable scores are a consequence of the focusing on the process of knowledge derivation instead of (implicitly) on the completeness of the experiment-space coverage. This way, the varying amounts of prior knowledge for different domains does not influence the consistency score as readily as it does the completeness and comprehension scores.

**OPTICS**

Apart from the methodological issues discussed in Chapter 5 on the Optics task, the conclusions from Chapter 5 resemble those from Chapter 3 and 4. Learners, even in the more complex Optics environment, are not focused on creating and testing hypotheses, but are mainly focused on constructing experiments. An analysis of learners' think aloud protocols showed that they do learn, however, that the learned knowledge does not necessarily overlap with knowledge tested in the post-test.

This is mainly caused by learners not knowing what they are looking for, both in terms of the final theory they are supposed to find (e.g., how the dependent variable is structured), nor in terms of which (independent) variables to use in explaining the behavior of the system. Again, learner's performance depend on their background knowledge (e.g., levels and variables to test) and a satisficing principle as testing all the combinations of levels and variables is pragmatically impossible.

**TASK GENERAL OBSERVATIONS**

Regardless the differences in the tasks described above, a number of observations can be made in each of these tasks, indicating that these observations are relatively task-independent:
Discovery skills used by learners in inductive learning tasks are often less complex than envisioned. That is, instead of extensive hypothesis-driven behavior, the behavior shown by learners in the studies in this thesis can to a large extent be described by a more simple algorithm: (1) think about something that can have an effect, (2) construct an experiment for the different situations in which that effect occurs, and (3) induce the existence of that effect based on the outcomes of these experiments. (See Chapter 3, and the rationale of the model presented in Chapter 2.)

Which variables are tested, and what levels are chosen by the learner for that variables depends on the task properties and prior knowledge, often taking the form of assumptions about the effect of variables. Moreover, the outcomes of experiments are interpreted by the learners in terms of this prior knowledge. If the newly discovered effects are unexpected, the learner is likely to engage in further examination of that variable. (See Chapter 2, 3 & 4.)

If no distinct stop criterion is given and the learner has no way to know whether the discovered effects are the ones searched for, a satisfying principle determines when to stop. That is, when new experiments are unlikely to uncover new results, the costs of conducting these experiments becomes higher than the profit associated with knowing these new results. In those situations, it becomes a rational decision to stop experimenting. (See Chapter 4 & 5.)

Based on the above, inductive learning in the tasks described in this thesis can be described as a process utilizing a relative straightforward experiment-construction strategy, which shows a pronounced influence of prior knowledge, and constrained by a stop-criterion based on bounded rationality.

THREE FACTORED EXPLANATION FOR INDUCTIVE LEARNING BEHAVIOR

The above discussion focused on the overlap between the three tasks discussed in this thesis along four dimensions. These dimension can also be generalized into three more general factors that are related to the learners' behavior. These factors
explain learners’ inductive learning behavior at a global level, and also explain why learners’ performance is often below the expected levels.

**Simplicity** Inductive learners appear to strive for an as simple as possible explanation for the observed behavior. Their behavior is probably best described by the principles of bounded rationality (Simon, 1957). Instead of performing an exhaustive search of the task’s experiment-space, learners weigh the possible gain of conducting more experiments against the associated costs. Even in relatively simple and well-structured tasks, the amount of effort associated with conducting all experiments prevents learners from conducting all experiments, making the learners resort to satisficing. Presumably, the perceived possible increase in knowledge does not countervail the costs of conducting more experiments. Another explanation for this effect is related to the law of diminishing returns. Initially, each correctly constructed experiment enables a learner in the Peter-task to derive a new main effect simply by comparing the previous experiment with the just constructed experiment. However, after conducting all the experiments associated with the main effects, a learner has to compare three already conducted experiments with a newly created experiment to derive a first order interaction. Moreover, if a learner assumes that simpler effects occur more often than more complex effects, searching for complex effects is more expensive because on average more experiments have to be constructed and compared per discovered effect. Therefore, the more a learner discovered, the smaller the chances are of discovering new information given the same investment.

Another aspect related to simplicity is that learners have a preference for simple hypotheses. If a simple hypothesis appears to predict the behavior of the task relatively well, learners are likely to stick to that hypothesis. Although this can be seen as an example of confirmation bias (Klayman & Ha, 1987) or as a preference for parsimony, this can also be accounted for on the basis of bounded rationality. As complex hypotheses tend to be both more complex to construct and to test, keeping an hypothesis as simple as possible decreases the chance to get entangled in complex hypothesis testing.

**Guiding knowledge** Apart from the aim for simplicity, learners’ behavior is also influenced by the application of prior knowledge. The most straightforward example of this notion is the importance of the availability of the weight and
distance concepts in the balance scale model presented in Chapter 2. A similar effect is present in models of Peter-task behavior, presented in Chapter 3. Besides testing the five variables presented in the interface, the learners only test for effects that seem probable or plausible on the basis of their prior knowledge about the domain. Given the overall complexity of the Optics task it is harder to pinpoint the exact contribution of prior knowledge. However, even in the relative short protocols presented in Chapter 5, both positive and negative effects of prior knowledge are clearly visible.

Thus, in all of these tasks, learners' behavior is actively guided by their knowledge about a domain. Not only in terms of their procedural knowledge that describes how to conduct correct experiments, but also in terms of declarative knowledge associated with the domain of the task.

An example of this is the difference in performance in the two tasks presented in Chapter 3. In the Peter-bikes-to-school task, learners are not told about the possibility of interactions between variables. Only two of the fifteen learners correctly discover and report the interaction during the post-test. In the Peter-goes-shopping task, the instruction contains information about the existence of an interaction, without making explicit how this interaction should be tested. Five of the fifteen learners in this task correctly report the interaction. Moreover, when these learners are given the Peter-bikes-to-school task directly after being tested in the Peter-goes-shopping task, thus after their concept of interaction has been “primed”, 9 out of 15 discover the interaction in this task whereas without being primed only 2 learners discover the interaction (Schoutsen, 1999).

More formally, given the aim for simplicity, the knowledge a learner has about the task is used in a process that adheres to the satisficing principle and limits the effective hypothesis and experiment search space to the regions of the total space in which the learner expects an effect (cf. the learner search space, Van Joolingen & De Jong, 1997).

**Salient Discrepancies** The “guiding knowledge” not only has a direct effect on the inductive learning task in that it determines which effects are researched, but also plays a role in the evaluation of discovered effects. In both the Peter-tasks and in Optics, two types of reactions have been observed after encountering a discrepancy between prior knowledge or hypotheses and data. In most of
the situations, learners just acknowledged the newfound effect, and revised their earlier beliefs. However, in some situations, learners were struck by the discrepancy and engaged in further research, trying to figure out what caused this discrepancy. This elaboration caused by the salient discrepancy often led to improved knowledge about the domain (e.g., in the Peter-task, discovering the interaction or in the Optics-task discovering the effect of the virtual focus point). This way, prior knowledge about possible effects does not only guide inductive learning by guiding the focus on what is tested, but the saliency of the discrepancy also determines whether the learner discovers less shallow effects in the domain under study.

Note that the notion of saliency also plays an important role in the simulations in Chapter 2. By means of increasing the saliency of a yet unused feature, the learner becomes aware of this feature and incorporates it into the discovery process. Although saliency plays a somewhat different role in this task compared to the Peter and Optics tasks, in all three tasks the saliency of features or concepts is an important predictor of whether or not that feature or concept is included in the final knowledge of the task.

**CONCLUSIONS**

The analyses presented in this thesis show that the claimed generality of the SDDS theory (Klahr & Dunbar, 1988) is probably overstated (see for a similar argument Johnson & Krems, 2001). In simple inductive learning tasks, as often encountered in normal life, hypothesis construction does not play the all-important role as sketched in the SDDS theory. Instead, learners seem to “simply” test for the effect of variables. This tendency toward *simplicity* also plays a more general role in inductive learning. Learners tend to keep their representations of the task as simple as possible. The level of this simplicity is determined by what is deemed necessary to test by the interface combined with *guiding knowledge* that states which variables and what type of relations are tested. If during the initiated tests *salient discrepancies* are found between the experimental outcomes and the assumed effect, elaborate on the found results. Because of this, these discrepancies can guide the learner toward a more complex set of hypothesis than initially constructed on the basis of the simplicity and guiding knowledge factors.
IMPLICATIONS

As was shown in all three tasks, learners have a difficult time discovering complex relations if they do not have guiding knowledge. That is, without explicit help, learners do not discover the multiplication rule in the balance scale task, seldom discover the interactions in the Peter-task, and do not easily discover the important landmarks in the Optics task. This clearly illustrates that the emphasis on “self-discovery” in modern curricula of secondary schools can only lead to satisfying results if the students are actively guided in their discovery process and if the tasks and domains to which discovery learning is applied are carefully chosen. First, students need to have relevant prior knowledge, so that they know what to look for. Second, they have to be aware that it is not necessary to cover the complete experiment-space. However, they also have to be aware that this only holds if they select their experiments from all parts of the experiment space (e.g., using heuristics like “test the extreme values” or “when you think that you’re ready, do some more random experiments to see if your predictions hold”). And third, given the important role of discrepancies in discovering the more complex relations in all three tasks, students need to be aware of the importance of discrepancies between their own hypotheses and assumptions and the discovered effects. Only if these three conditions are met, discovery learning might be a useful addition to more traditional educational methods.
CHAPTE RR 7:

SAMENVATTING

Als mensen “inductief leren”, leren zij door van een aantal voorbeelden algemene regels af te leiden. Deze regels zijn vaak algemeen van aard; ze bestrijken een groter deel van de probleem ruimte dan daadwerkelijk is waargenomen. De voorbeelden die ten grondslag liggen aan de regels kunnen zowel door de leerder zelf zijn gegenereerd als door een externe entiteit worden aangeboden. Zeker in de situaties waarbij de voorbeelden door de leerder moeten worden gegenereerd, is het niet ongebruikelijk dat inductief leren gelijk wordt gesteld met “in de wetenschap” gebruikelijke methodes van onderzoek. Op basis van het actieve karakter van dit leren, verbonden met het “zelf onderzoeken en ontdekken”, is geclaimd dat dit een goede manier is om robuste kennis op te doen. Echter, experimentele studies lijken dit niet te bevestigen. Leerders doen slechts weinig kennis op tijdens inductief leer sessies. Een van de vragen onderliggend aan dit proefschrift is dan ook wat de oorzaak is van deze relatief teleurstellende inductief leer resultaten. Hiertoe wordt het gedrag van leerders in drie verschillende taken bestudeerd. Een belangrijk aspect hierbij is hoe het gedrag van de leerder gemeten en beoordeeld wordt. In plaats van de gebruikelijke focus op kenmerken van leerders die verklaren waarom de effectiviteit van leren varieert, ligt de focus in dit proefschrift vooral op de interactie tussen het gedrag en de kennis van de leerder en hoe de beoordeling van het gedrag tot stand komt.

Echter, voordat inductief leren in volwassen besproken wordt, gaat dit proefschrift eerst in op het leren tijdens ontwikkeling. In tegenstelling tot leren later op school of tijdens psychologische experimenten in standaard leer-omgevingen is het uitzonderlijk als jonge kinderen correctheids-feedback krijgen op hun cognitief gedrag. Hoofdstuk 2 beschrijft een computationeel model dat het gedrag van kinderen op de balans-taak simuleert. In een balans-taak krijgt een kind een balans te zien waarop
op verschillende afstanden zowel links als rechts van het scharnierenpunt gewichten kunnen worden geplaatst. De opdracht is om te voorspellen naar welke kant de balans zal schaarnieren. Het ontwikkelde model is, in tegenstelling tot alle bestaande modellen van balans-taak gedrag, in staat om het zelfde type leergedrag te laten zien als kinderen, ook als er geen feedback wordt gegeven aan kind of model over de correctheid van het antwoord. Daarnaast laat dit model zien dat uit het samenspel van één simpele regel ("zoek de verschillen") en een aantal onderbouwde aannames de verschillende stadia van gedrag verklaard kunnen worden.

De overige hoofdstukken hebben betrekking op meer traditionele inductief leer taken. In de Hoofdstukken 3 en 4 zijn Peter-taken onderwerp van studie. De opdracht aan een leerder in een Peter-taak is uit te vinden hoe de nominale waardes van een set gegeven onafhankelijke variabelen de (semi-)continue uitkomst op een afhankelijke variabele bepalen. Hiertoe kan een leerder zelf experimenten genereren door een waarde te kiezen voor ieder van de gespecificeerde variabelen, waarna de waarde van de afhankelijke variabele opgevraagd kan worden. Dit is een van de taken die zich meer leent voor de vergelijking met hoe in de wetenschap onderzoek wordt gedaan. Echter, de op computationele modellen gebaseerde analyses die worden besproken in Hoofdstuk 3 tonen een belangrijk verschil aan tussen theorieën over wetenschappelijk onderzoek en het gedrag van leerders. Een centraal element in theorieën over wetenschappelijk onderzoek is "de hypothese". Volgens deze theorieën zou "goede wetenschap" gefocusséerd moeten zijn op het construeren en testen van hypotheses. Op basis van een vergelijking tussen de modellen in Hoofdstuk 3 en het gedrag van leerders in de Peter-taak wordt echter beargumenteerd dat leerders in deze taak hun gedrag nauwelijks laten sturen door hypotheses. In plaats daarvan laten zij zich sturen door de layout van de taak en hun voorkennis of aannames over de variabelen in de taak. Een ander aspect dat naar voren komt bij de bespreking van de modellen is dat de mate van succes zoals gemeten in traditionele tests sterk afhankelijk is van de voorkennis en assumpties van die leerder. Niet omdat de test direct deze voorkennis toetst, maar omdat de voorkennis de leerder kan leiden naar anders onontdekt gebleven kennis.

Deze afhankelijkheid van voorkennis is de focus van Hoofdstuk 4. In dit hoofdstuk wordt beargumenteerd dat de standaard tests die gebruikt worden om inductief leren te meten niet voldoende zijn. Als men geïnteresseerd is in de kwaliteit van
Samenvatting

het inductief leren, dan is het noodzakelijk taak-afhankelijke aspecten uit te sluiten. Echter, gezien de sterke invloed van voorkennis op de gemeten scores is een maat die puur de nieuw opgedane kennis meet niet noodzakelijkerwijs een goede maat voor de kwaliteit van het ontdekgedrag. Als alternatief voor een puur op een eindtest gebaseerde score wordt in Hoofdstuk 4 een maat besproken die de consistentie tussen het gedrag van een leerder en de score op een eindtest meet. Deze maat wordt onderbouwd door twee verschillende Peter-taken te vergelijken, waarbij wordt aangetoond dat de op consistentie gebaseerde score minder beïnvloed wordt door de veranderingen in de taak-setting.

In de Peter-taak worden zowel de onafhankelijke als de afhankelijke variabelen expliciet gemaakt voor de leerders. Hoewel zij de experimenten zelf moeten genereren, hoeven zij daardoor niet meer uit te zoeken welke variabelen mogelijkerwijs van belang zijn. Daarnaast is de Peter-taak uitzonderlijk in de discreetheid van de variabelen: slechts in weinig gevallen is de werkelijkheid zo goed op te delen in nominale waardes als in de Peter-taak.

Optics, de taak die besproken wordt in Hoofdstuk 5, is speciaal ontworpen om inductief leren te meten in een setting die gelijkwaardiger is aan de complexe werkelijkheid. In de standaard Optics simulatie setting krijgt de leerder een lege werkbank aangeboden waarop lenzen en lampen geplaatst kunnen worden. Door het verplaatsen van objecten of het veranderen van eigenschappen (bijvoorbeeld de hoek die de lichtstraal maakt met de optische as) kan de leerder zelf experimenten genereren. Een belangrijk aspect hierbij is dat tijdens het manipuleren de effecten hiervan direct te zien zijn. Er is dus sprake van continue in plaats van discrete experimentatie. Echter, de inductief leer resultaten in deze taak worden tegen. Zonder sturing of imperking kwamen de scores op post-tests nauwelijks uit boven pre-test scores. Door het minder gestureerde karakter van Optics zijn de methodes zoals besproken in Hoofdstuk 4 echter niet direct toepasbaar op het analyseren van inductief leer-gedrag in Optics. Om deze analyses wel mogelijk te maken, is een kwalitatief redeneer model opgesteld dat de continue experimentatie discreet analyseerbaar maakt. Door middel van de analyse op basis van het kwalitatief redeneer model wordt ook in deze taak aangetoond dat leerders wel degelijk adequate inductief leer strategieën beheersen. Echter, doordat het onduidelijk is voor leerders wat zij worden geacht te ontdekken en doordat een belangrijke variabele niet zonder meer te identificeren valt, besluiten
veel leerders te stoppen voordat alle kennis verzameld is.

CHAPTER 8:

SUMMARY

When people engage in “inductive learning” they try to induce general rules from a set of specific examples. These rules are general in the sense that they cover a larger portion of the problem space than which has actually been explored. The examples that are at the basis of these rules can either be generated by the learner or they can be provided by an external entity. If the examples are generated by the learner, inductive learning is often referred to as a “scientific” discovery process, in analogy to the way scientific research is being conducted. Due to the active character of the learning process, combined with “exploration and discovering”, it has been claimed that this is a good way of acquiring robust knowledge in a new domain. However, experimental studies do not seem to support this claim. Learners learn only very little during inductive learning sessions. One of the questions that this thesis is based on is about the reason for these relatively disappointing results. Three different tasks and the respective behavior of learners have been studied to answer this question. An important aspect is how the learner's behavior is measured. Instead of focusing on properties of the learners that explain the variance in learning results, as is often done, this thesis focuses on the interaction between the behavior and knowledge of the learner and the aspects on which the behavior is measured.

Before discussing inductive learning in adults, a developmental study is presented that is related to inductive learning. In contrast to learning in a school-setting or during most psychological experiments, it is uncommon for young children to get correctness feedback on their cognitive behavior. Chapter 2 describes a computational model that simulates children’s behavior on the balance scale task. In a balance-scale setting, a child is presented a balance scale on which weights are placed left and right.
of the fulcrum. The child has to predict to which side the balance will tilt. The model presented in Chapter 2 can, in contrast to all existing models of balance scale behavior, explain the behavior of children, both in situations with and without feedback. This model also shows that an interaction between a simple rule ("search for differences") and a number of psychologically grounded assumptions is sufficient to explain the observed behavior.

The other chapters are focused on more traditional inductive learning tasks. Chapter 3 and 4 focus on the Peter-task. A learner in a Peter-task has to discover how the nominal independent variables determine the (semi-)continuous outcome on a dependent variable. A learner has to select a value for each of the independent variables to construct an experiment, after which the outcome on the dependent variable is presented. This is one of the tasks that suit the comparison between inductive learning and "scientific research". However, the computational models presented in Chapter 3 illustrate an important difference between the theories about the scientific method and the behavior of learners. The central topic in the scientific method is the hypotheses; "good science" is focussed on the testing and revision of hypotheses. It is argued, based on the comparison between the models in Chapter 3 and the observed behavior, that the behavior of learners in the Peter-task is not guided by hypotheses. Instead, their behavior is guided by the layout of the task and their prior knowledge of and assumptions about the variables in the task. Another aspect discussed is that the amount of success as measured in the traditional tests is largely dependent the prior knowledge and assumptions of the learner. Not because this prior knowledge is tested during the post-test, but because the prior knowledge can guide the learner to otherwise undiscovered knowledge.

This dependence on prior knowledge is the focus of Chapter 4. In this chapter, it is argued that the standard tests for measuring inductive learning are insufficient. If one is interested in the quality of inductive learning, it is necessary to exclude any task-specific influences. However, a test that solely measures the new gained knowledge is not appropriate for measuring the quality of inductive learning skills given the influence of prior knowledge on that test. Chapter 4 proposes a consistency measure as an alternative to knowledge-based tests that expresses the consistency between a learner's behavior and the post-test scores. This consistency measure is grounded by comparing two different Peter-tasks, showing that the new score is less influenced by
changes in the task-setting.

Both the dependent and the independent variables are made explicit to the learner in the Peter-task. Although learners have to construct their experiments, they are not required to figure out which variables might be relevant. Besides, the levels of the variables in the Peter-task are discreet; only in very few cases can the real-world be categorized in such clear-cut nominal values.

Optics, the task discussed in Chapter 5, was especially designed to measure inductive learning in a setting more close to the complex world as encountered outside the psychological laboratory. The learner is presented an empty workbench on which the learner can place lamps and lenses. Learners can generate experiments by moving the objects and changing the angle of the light beam originating from the lamps. An important feature of Optics is that the simulation is updated during all manipulations, yielding continuous experimenting instead of the discrete experimentation in the Peter-task. The learning results in this task are often disappointing. The scores on the post-test were comparable to those on the pre-test in the settings without guidance. Because of the less structured character of the Optics simulation the analysis methods as discussed in Chapter 4 are not directly applicable to the analysis of behavior in Optics. A qualitative reasoning model was constructed to be able to analyse the continuous experimenting in a discrete fashion. By analyzing the data using the qualitative reasoning model it is shown that learners to possess adequate inductive learning strategies, as was shown for the Peter-task. However, because it is not clear to learners what they have to discover and because one of the important variables is not easily identified, a lot of learners decide to stop experimenting before all effects are discovered.

Chapter 6 concludes this thesis by discussing the relations between the three tasks that are analyzed in the thesis. Based on the previous chapters this chapter introduces three factors that determine the behavior of learners and their success during an inductive learning session. According to the rationale behind the first factor, learners try to keep the derived knowledge as simple as possible. The second factor is involved with the importance of prior knowledge in the guidance of the inductive learning process. And “saliency”, the third factor, is related to the importance of the saliency of an unexpected outcome.
If these factors are taken into account, the behavior of inductive learners is not as bad as the meager post-test results seem to suggest.
REFERENCES

ACT-R 5.0 (n.d.) Retrieved October 5th, 2002, from “ACT-R 5.0 Beta” Website: http://act.psy.cmu.edu/ACT-R_5.0/

ACT-R 5.0, Unit 9: Production Rule Learning (n.d.) Retrieved January 5th, 2002, from “ACT-R 5.0 Beta” Website: http://act.psy.cmu.edu/ACT-R_5.0/unit9/

Thinking Aloud (n.d.) Retrieved October 31, 2001, from “Usor, A Collection of User Oriented Methods” Website: http://sunrise.nada.kth.se/usor/jml.cgi/Methods/thinking.jml


References


References


Dit proefschrift is het vierde en laatste proefschrift dat tot stand is gekomen binnen het NWO aandachtsgebied Inductief Leren. Voor mij is Casper Hulshof gepromoveerd op het proefschrift *Discovery of Ideas and Ideas about Discovery. The Influence of Prior Knowledge on Scientific Discovery Learning in Computer-Based Simulations*, Pascal Wilhelm op het proefschrift *Knowledge, Skills and Strategies in Self-directed Inductive Learning* en Frans Prins op het proefschrift *Metacognitive Influences on Inductive Learning*. 