COGNITION AND INSTRUCTION, 1986, 3(1), 63-86

Learning to Understand the Balance Beam

Pamela Thibodeau Hardiman, Alexander Pollatsek, and Arnold D. Well
University of Massachusetts

Twenty-two university students who did not initially know the quantitative rule for predicting whether a configuration of weights placed on a balance beam would cause the beam to balance, tip left, or tip right were asked to induce the rule in a training procedure adapted from Siegler (1976). For each of a series of balance beam problems, subjects predicted the action of the beam and explained how they arrived at their prediction. Protocols revealed that although all subjects realized early on that both weight and distance were relevant to their predictions, they used a variety of heuristics prior to inducing the correct quantitative rule. These heuristics included instance-based reasoning, qualitative estimation of distance, and the use of quantitative rules of limited generality. The common use of instance-based reasoning suggests that learning to understand the balance beam cannot be described completely in terms of a simple rule acquisition theory. Also, the variability in the use of heuristics across subjects suggests that no simple theory that depicts subjects as linearly progressing through a hierarchy of levels can adequately describe the development of balance understanding.

In this article, we are concerned with the way in which people develop an understanding of physical concepts that have the following properties: (a) Two separable variables are involved; (b) those variables must be measured or quantified in some way; and (c) the measurements may be combined according to some rule, yielding a third quantity or construct that allows prediction of what will happen for any combination of the two variables. Examples of such concepts include density, the size of shadows, and torque.

Specifically, we investigated how people combine information about number of weights and their distance from the fulcrum to predict whether a balance beam will balance, tip left, or tip right. A balance beam consists of a bar placed on top of a fulcrum or balance point. Weights may be placed on both

Requests for reprints should be sent to Pamela Thibodeau Hardiman, Department of Physics, Hasbrouck Hall, University of Massachusetts, Amherst, MA 01003.

sides of the fulcrum and the effectiveness of a weight in causing the beam to tip is determined by the product of the weight (w) and its distance from the fulcrum (d), a construct called the torque associated with the weight. If the total torque (i.e., $\Sigma w_i d_i$) associated with the weights on each side of the beam is the same, the beam will balance; otherwise, the beam will tip to the side with the greater torque. Because of its form, this rule will subsequently be referred to as the *product-moment rule.*

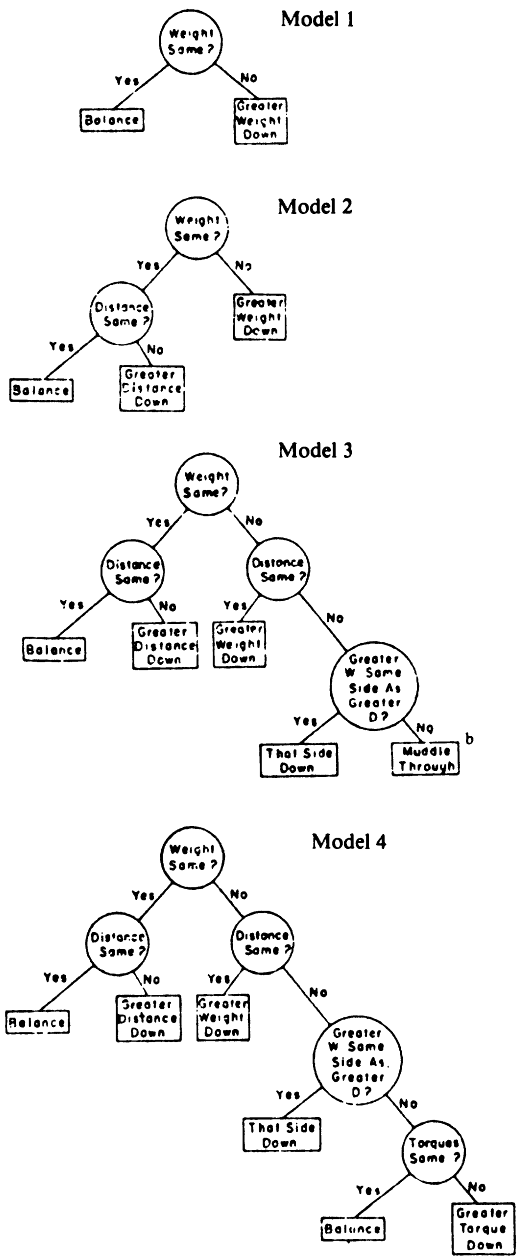
There are a number of reasons why balancing is an important and rich domain in which to study how subjects learn to combine information about variables. Most people have some understanding of the factors that determine whether balancing will occur. Even fairly young children can often identify weight and distance from the fulcrum as the critical variables, and provide reasonable intuitive explanations of why a set of weights will or will not balance on the balance beam. On the other hand, relatively few adults can specify a rule that will allow them to predict what will happen in any given situation. In fact, over several experiments only about 20% of adults have produced responses to balance beam problems consistent with the product-moment rule (Jackson, 1965; Lovell, 1961; Siegler, 1976). Even when provided with specific experiences intended to promote understanding of the concept of balancing, adults do not easily derive the product-moment rule. It is the period between being able to identify the relevant variables and being able to combine them so as to make correct predictions that is of major interest to us.

Why is it so difficult to generate the product-moment rule? In fact, the rule is easy to use: Siegler (1976) taught three 10-year-old children the product-moment rule, and they were subsequently able to use it successfully. It is clear, however, that the rule will be induced only if the relevant features or dimensions of the problem are identified and combined appropriately.

There have been two major descriptions of the phases of understanding the balance beam: Inhelder and Piaget's (1958) stage theory, consisting of three stages, each divided into two substages; and Siegler's (1976; Klahr & Siegler, 1978) hierarchical rule models. Both descriptive systems are presented in Figure 1.

Inhelder and Piaget (1958) presented two types of tasks. In the first task, they employed a balance beam that had holes at equal intervals on both sides of the fulcrum (28 holes on each side) and weights of differing sizes that could be hung at various distances from the fulcrum. In the second task, childen were presented with a balance with no holes in the crossbar, and instead of weights there was a basket on each side into which dolls could be placed. No units were marked along the crossbar. Subjects, who ranged in age from 3 to about 14 years, were to told play with the balance beam to find out how it worked. Although in the first task it was possible to hang weights at more than one location on each side of the fulcrum, apparently even the most ad-

Siegler's Classification[a]



Inhelder and Piaget's Classification

Stage 1A: Subjects fail to distinguish their own actions from external processes (e.g., the subject will push the beam so that it is level and expect it to remain that way).

Stage 1B: Subjects realize that weight is needed on both sides of the fulcrum to achieve balance, but there is as yet no systematic correspondence between weight and distance.

Stage 2A: Subjects achieve balance by making weight and distance both symmetrical. Subjects discover by trial-and-error that there is equilibrium between a smaller weight at a large distance from the fulcrum and a greater weight at a small distance but do not draw out general consequences.

Stage 2B: Subjects develop qualitative understanding of the relationship between weight and distance.

Stage 3A: Subjects start to discover the quantitative law for balancing.[c] It takes the form of the proposition $W / W' = L' / L$, where W and W' are two unequal weights and L and L' are the distances from the fulcrum at which they are placed.

Stage 3B: Subjects search for a causal explanation.

[a]After Klahr and Siegler (1978).

[b]"Muddle through" means "guess."

[c]At least for the special case in which weights are placed at only one distance on each side of the fulcrum.

Figure 1 Siegler's and Inhelder and Piaget's classifications.

vanced subjects did not do so. Therefore, the rule $w_1 / w_2 = d_2 / d_1$ (henceforth referred to as the *ratio rule*), where w_1 and w_2 are the amounts of weight on each side of the fulcrum and d_1 and d_2 the corresponding distances from the fulcrum, would always be sufficient to predict when balancing would occur, and the general form of the product-moment rule was not needed.

Examination of Inhelder and Piaget's stages in Figure 1 suggest an increasingly systematic approach to understanding the influence of the variables of weight and distance. However, the usefulness of the description is limited by the vagueness of some of the terms employed, such as *qualitative understanding*. In addition, it is not clear what is necessary to advance from one stage to the next. Finally, as we have indicated earlier, Inhelder and Piaget discussed the ratio rule but not the most general form of the product-moment rule.

Siegler (1976) used a balance beam that had four pegs on each side at equally spaced distances from the fulcrum and equally sized metal weights that had holes in their middles so that they could fit over the pegs. Subjects, who ranged from kindergartners to 12th graders, had their understanding of the balance beam assessed after having participated in one of three conditions. In the a priori (i.e., control) condition, subjects were not given any experience with the beam before being tested for understanding. In the experimentation condition, subjects were told that there were rules by which they could predict the action of the balance beam and that they should "experiment" with the beam and the weights and try to learn how the beam worked. In the observation condition, subjects were also told that a rule existed and were presented with a predetermined series of 36 problems whose outcomes they could observe.

Understanding of the balance beam was assessed in a posttest by presenting each subject with a series of 30 problems that would potentially distinguish different levels of knowledge. For each configuration of weights, subjects were asked what would happen if the balance beam (which was kept level by two wooden blocks that were placed beneath the arms of the balance) were released. No feedback concerning the action of the balance beam was provided. Posttest performance was approximately the same for subjects in the a priori, experimentation, and observation groups, suggesting that not much learning took place in the experimentation and observation conditions.

Siegler proposed that performance on the balance beam can be described by the set of four hierarchical models presented in Figure 1, and further, that an individual progresses developmentally from model to model in an invariant order. Siegler classified the responses of 107 of the 120 children as conforming to the predictions made by one of the four models and concluded that the study provided considerable support for the descriptive accuracy of the models. The responses of most (23 of 30) 5- and 6-year-olds were predicted by Model 1, while the responses of most (48 of 90) of the older children were predicted by Model 3. Notably, the responses of only 5 of the 30 children in the oldest group (16- and 17-year-olds) were predicted by Model 4.

Siegler's (1976, 1978; Klahr & Siegler, 1978) rule-based set of developmental models has several advantages over Inhelder and Piaget's (1958) stage theory. In particular, the models present a greatly simplified interpretation of complex behavior. The models make testable claims concerning which balance beam problems will be answered correctly by a subject performing at the level of one of the four models. In addition, specification of the models in terms of production systems (Klahr & Siegler, 1978) makes it possible to state how, by modifying existing productions or adding new ones, more advanced models are created. The clarity and testability of these models allows questions relevant to instruction, such as what kinds of instruction are effective in producing movement to more complex models and whether people of different ages are differentially responsive to instruction, to be posed.

Although the simplicity of Siegler's hierarchical models make them an attractive system for predicting the behavior of people interacting with a balance beam, several potential difficulties arise when they are viewed as models of the reasoning process, rather than merely as predictors of behavior. One problem is that, as Strauss and Levin (1981) pointed out, "the rules of the rule system are the outcome of an interaction between task variables and an overarching cognitive system that attempts to deal with them" (1981, p. 76), so that the relative simplicity of Siegler's rules as compared to Inhelder and Piaget's descriptive system may in part reflect the more structured nature of the tasks that Siegler employed. In addition, we argue that there are reasons to believe that systematic predictions are based on considerations other than those suggested by the decision trees that make up Siegler's models, and for that reason these decision trees are insufficient models of the reasoning process. There are three general areas of concern: (a) what is involved in the change from Model 3 to Model 4, (b) how distance is encoded (i.e., whether distance is dealt with as an ordinal variable or encoded numerically) at various levels in the hierarchy, and (c) the limitations in the knowledge the models allow to be used in making a decision about a balance problem.

THE TRANSITION FROM MODEL 3 TO MODEL 4

The hierarchical rule models are relatively weak in specifying how transitions between models occur, largely because Model 3 is poorly specified. Although the production system formulation of the models presented by Klahr and Siegler (1978) indicates what alterations must be made to Model 3 to yield Model 4, the processes by which the changes occur are not obvious. Model 4 is derived by altering Model 3 such that when a conflict problem is encountered, predictions are no longer based on "muddling through" (i.e., randomly guessing). Instead, they are based on computing and comparing the torques for each arm of the balance beam (i.e., predictions are based on the product-moment rule). A strict interpretation of Siegler's rule hierarchy as a developmental model of reasoning would suggest that the transformation from a system dealing with conflict problems in a random way (Model 3) to one dealing with them correctly (Model 4) occurs in a single step. There seems reason to believe that at least some subjects pass through an intermediate stage in which they generate rules that incorporate both weight and distance but are of limited generality. In Inhelder and Piaget's classification, the quantitative rule arrived at in Stage 3 is the ratio rule, not the more general product-moment rule. Moreover, Klahr and Siegler's (1978) detailed analysis of a single subject and description of Model 3 explanations advanced by subjects seem to indicate that the use of limited rules is probably quite common.

Siegler (1976) classified subjects' responses as conforming to Model 3 if there were (a) fewer than 26 correct responses on the 30 posttest problems, (b) at least 10 correct responses on the 12 nonconflict problems (in which either weight or distance or both were the same on both sides of the fulcrum or else the greater weight was on the same side as the greater distance), and (c) more than four departures from complete reliance on the weight dimension on the 18 conflict problems. However, these criteria could have been met by subjects employing a variety of rules and strategies (cf. Wilkening & Anderson, 1982) such as using the ratio rule or special cases of it (e.g., 2:1), or even comparing the results of adding weight and distance on each side of the fulcrum rather than multiplying them (which would correctly predict that configurations like 0100/3000 would balance but would not work in general). It seems clear that Klahr and Siegler themselves considered Model 3 to be an umbrella classification for a host of more specific strategies that are inconsistently adopted, although they did not incorporate these strategies into their description of Model 3.

An important question, therefore, is whether subjects generate rules, such as the ratio rule, that combine weight and distance information but do not apply in all situations, before they learn the more general product-moment rule.

ENCODING OF DISTANCE

A second area of concern is the manner in which distance is encoded. The use of distance first occurs in Model 2, in which distance is considered if the weights are the same on both sides of the fulcrum. However, the only judgments about distance necessary for Model 2 or Model 3 are ordinal, that is, whether distance on one side of the fulcrum is less than, equal to, or greater than the distance on the other. Only for Model 4 must both weight and distance be encoded in a quantitative or numerical fashion so that torque may be calculated. It is not clear whether subjects performing at the level of Model 2 or 3 encode distance numerically, as Siegler's (1976) research would seem to imply, make crude perceptual judgments, or simply make ordinal decisions about distance. If subjects do not encode distance numerically, they will have difficulty generating the product-moment rule, and this may result in the use of alternative strategies.

The manner in which subjects encode distance is therefore of critical importance in building a model of how they reason before arriving at the product-moment rule. The apparatus used by Siegler (1976) does not particularly lend itself to the investigation of this issue, since the four pegs on each side of the fulcrum at which weights could be placed are very salient. In the present study, we attempted to gain insight into how subjects encoded distance in a situation in which a measurement scale was provided but was not so obvious.

USE OF INFORMATION ABOUT PREVIOUS PROBLEMS

According to Siegler's rule-based models, a subject confronted with a balance problem will go through a decision tree but will not use information about either specific problems or general types of problems encountered previously. It seems plausible, however, that subjects will compare a given configuration of weights with a pattern observed earlier at least some of the time. In fact, some of the experience with the balance beam that Siegler provided his subjects would seem to encourage such comparisons. In his observation condition, subjects were presented with four sets of five problems each, in which problem $n + 1$ differed from problem n by the addition or removal of a single weight. In such a series of problems, there are several points at which the correct outcome could be predicted on a logical basis merely by considering the change from the previous problem. For example, if a single weight is added to one arm of a configuration that previously balanced, the beam must tip to that side.

One basis for building a rule-based set of models such as Siegler's is the assumption that because the final state of learning can be described in terms of a rule, the intermediate stages are also best described in terms of rules. There is reason to question this assumption. Brooks (1978) suggested that when a rule for categorizing a set of exemplars exists but is sufficiently complex and difficult to induce, subjects may categorize new instances in terms of their similarity to previously observed instances rather than by generating and using a rule. Although the product-moment law is not as complex a rule for categorization as the examples used by Brooks, it is obviously difficult to induce. Balance beam problems might be considered to be members of one of three categories: (a) "balance" configurations, (b) "tip right" configurations, or (c) "tip left" configurations, and the decision about a balance beam problem conceived of as a problem in categorization. Even when told that a rule exists that will allow correct predictions for all balance problems, subjects may, at least before the rule has been generated, make decisions about balance problems by comparing them with representations of previously observed problems that are exemplars of these categories. At the very least, Brooks's findings suggest the possibility that predictions made before the product-moment rule has been learned may not all be based on the use of simple rules.

In summary, although the rule-based models proposed by Siegler (1976) and Klahr and Siegler (1978) have appeal as a system within which to describe the predictions people make when given balance problems, it appears there are reasons to believe that they are not adequate models of the reasoning process. In the present study an attempt was made to determine whether, when provided with a series of balance problems and given the task of inducing the product-moment rule, people (a) generate rules of limited generality that involve both weight and distance before they learn the general product-moment rule, (b) encode distance numerically at different levels of performance, and (c) base judgments about balance problems on specific information about previously encountered problems or classes of problems.

The study consisted of two phases. In the pretest phase, a paper-and-pencil test was used to identify subjects who had not yet learned the product-moment rule. In the training phase, 22 of these subjects were presented with a series of balance problems using wooden blocks and a balance beam. Although the set of balance problems was modeled after that used in Siegler's (1976) observation condition, the procedure differed in several important ways. In Siegler's study, the observation condition consisted of 36 trials, on each of which the experimenter placed weights on the balance beam, removed the wooden blocks holding the beam level, and allowed the subjects 10 seconds to observe the outcome. In the present study, subjects were told to predict the outcome for each balance problem before being allowed to observe the outcome, were asked to think aloud while performing the task and, if possible, to provide justification for their prediction. In addition, the training phase continued until subjects gave answers consistent with the product-moment rule.

METHOD

Pretest Phase

Subjects. Forty-eight students (30 women and 18 men) enrolled in psychology classes at the University of Massachusetts received bonus credit for participation in the pretest phase of the study. Age ranged from 17 years to 36 years, with a mean of 20.5 years.

Problems. The pretest consisted of 12 schematic line drawings of a balance beam presented with various configurations of weights. These problems varied in both type and difficulty. There were three "simple" problems that did not require the product-moment rule for correct predictions because either (a) weight or distance or both were equal on both sides of the fulcrum (e.g., 0300/0300. Note that the notation indicates the number of weights one, two, three, and four units of distance on either side of the fulcrum; in this case, there are three weights located three units of distance to the left of the fulcrum and three weights located two units of distance to the right of the fulcrum), or (b) the greater weight was associated with the greater distance (e.g., 0120/1100). The remaining nine problems represented "conflict" situations in which the greater weight was associated with the lesser distance. For three of these problems, the beam would have tipped to the side with greater weight (e.g., 0220/0002), and for three it would have tipped to the side with the greater distance (e.g., 1100/2200). The remaining three problems presented balance situations (e.g., 1100/1300). The problems employed depicted situations with weights placed at no more than two locations on either side of the fulcrum.

Procedure. The pretest was administered to groups of three or four subjects. Subjects were instructed to predict whether the balance beam presented in each problem would tip to the left, balance, or tip to the right. No time limit was imposed and feedback was not provided.

Analysis. All subjects who made three or more incorrect predictions were classified as nonbalancers. Thirty-six of the 48 subjects who took the pretest were classified as nonbalancers.

Training Phase

Subjects. Twenty-two of the 36 subjects classified as nonbalancers on the pretest (16 men and 6 women) were randomly chosen to participate in the training phase of the study. These subjects had a mean score of 6.40 correct answers on the pretest, with a standard deviation of 1.59.

Materials. The balance beam used in the training phase consisted of a flat, rigid aluminum bar balanced on a fixed fulcrum at its midpoint. Distance from the fulcrum was denoted by marks at regular intervals drawn on a lightweight acetate scale placed on top of the aluminum bar. The unit marks were approximately 5 cm apart, and half units were also indicated. It should be emphasized that in contrast to Siegler's (1976) study in which weights could be placed only on four pegs on each side of the fulcrum, in the present study, weights could be moved continuously along the surface of the beam. The weights were wooden cubes whose sides were approximately 3.5 cm.

Problems. The set of problems used was modeled after that used in Siegler's (1976) observation condition. Problems presented near the beginning of the session were relatively simple and became progressively more complex. The first few problems consisted of single blocks or stacks of blocks placed on each side of the fulcrum such that either the weight or distance (or both) was equal on both sides of the fulcrum. These problems were followed by three sequences of problems each of which began with two equal stacks of blocks, one placed closer to the fulcrum than the other. Blocks were added one at a time to the stack closer to the fulcrum until the beam balanced and then tipped in the opposite direction. Thus, subjects saw how problems with conflicting weight and distance could tip to either side or balance. After completing these three sequences of problems, the balance situations from each sequence were repeated. The final and most complex problems employed two stacks of blocks on one or both sides of the fulcrum. There were up to six blocks per stack.

Procedure. Before the session began, the subject was informed of and consented to the videotaping of the interview. Each subject was interviewed individually, while seated opposite the interviewer with the balance beam between them.

The subject was told that he or she would be given a series of balance problems. The instructions stated that the task for each individual problem was to predict whether the beam would tip to the left, balance, or tip to the right; and further, that the subject should attempt to discover the general rule which would allow correct prediction for all the problems. Subjects were asked to think aloud while performing the task and if possible to provide justifications for the responses they made.

For each problem the interviewer placed blocks on the balance beam while holding it level. After the subject made a prediction and commented on it, the beam was released so that the subject could determine whether the prediction was correct. The interviewer then continued with the next problem.

In the first part of the training phase, all subjects received the same problems in the same order. There was, however, some variation in the later problems, as subjects were allowed to request that a particular configuration of weights be set up at any time, although they were not specifically instructed to do so. In addition, if the subject stated an incorrect hypothesis about the balance beam, the interviewer attempted to set up special problems that contradicted the hypothesis. The criterion for concluding that the subject had induced the product-moment rule was five consecutive correct predictions on complex problems (i.e., conflict problems with more than one stack of weights on each side of the fulcrum).

If, after 40 problems, the subject had not made any hypotheses about general rules and did not appear to be counting the numbers of distance units, the interviewer prompted the subject to pay closer attention to distance. She did so by asking, "How far out is that pile of blocks?," when the subject made a vague comment that a particular stack of blocks was "closer in" or "farther out."

Analysis. Three types of information were of major interest: (a) evidence that subjects based their reasoning on rules of limited generality that involved both weight and distance before they learned the product-moment rule, (b) evidence that subjects either did or did not encode distance quantitatively at different levels of performance, and (c) evidence that subjects based predictions on information about previously encountered problems or classes of problems.

Subjects were considered to have used ratio reasoning if they stated the relationship between the ratios of weight and distance in the problem either before or after making their prediction. Verbal references to counting and pointing to the distance marks were considered to be positive evidence of numeric encoding of distance. In addition, evidence for the lack of numeric encoding of distance was noted. Indications of gross perceptual judgments such as leaning back to judge distance, comments such as "It's hard to tell whether they're the same," and questions such as "Are the lines on the beam important?" were considered to be evidence that people were not encoding distance numerically.

The coding of reasoning based on prior instances was somewhat more complicated, given the less explicit nature of the responses. Coding was done by two independent coders with disagreements resolved by a third, so that a datum had to be agreed upon by at least two of the three coders before being counted. Reasoning from particular problems and from particular classes of problems (i.e., those with a particular relationship between the ratios of weight and distance) were included in this classification. References to particular instances were broken down into problems in which a single change had been made to the previous problem and those in which two changes had been made to the previous problem. In order for an explanation to be classified as reasoning from an instance, two things had to be true. First, the explanation had to contain an explicit reference to a previous problem such as "it balanced before but now it doesn't because this weight is farther out" or a fairly clear implicit reference such as "this is farther out now and so it shouldn't balance." Second, the explanation had to make sense to the coder in terms of the problems that had actually preceded the explanation. A consequence of the latter restriction was that for most of the explanations classified as reasoning from an instance, the instance reasoned from had been presented only one or two problems back. The decision to restrict this classification to only those explanations in which the instance-based reasoning was clearly articulated and based on an accurate memory of a previously presented problem probably caused us to underestimate the occurrence of instance-based reasoning.

The reasoning was coded as logically correct or incorrect based on the relationship between the previous and the current problem and the subject's response. For example, if the subject noted that the beam was previously balanced and one block was added so that it must now tip to the side of the added block, he or she was logically correct. However, if the beam had been tipped to one side and one block was now added to the other side, the subject would be incorrect in reasoning only on the basis of that fact that the beam should now balance. References to classes of configurations previously encountered were also coded into logically correct and incorrect forms. Correct forms included (a) two sets of ratios that will balance if "added together" (e.g., 0024/0201 must balance because it is composed of 0004/0200 and 0020/0001, both of which balance), and (b) a configuration of weights that will not balance if it differs in a critical way from a ratio configuration that does balance (e.g., 0100/3100 will not balance because it differs by one block from 0100/3000, which balances). Incorrect predictions included those based on the similarity between the current configuration and one previously encountered (e.g., 1100/2200 is predicted to balance because it is similar to 2000/0400).

RESULTS

All 22 subjects were able to meet the criterion for learning the product-moment rule. The mean number of trials required to meet the criterion was 49.0, with a standard deviation of 15.4 (range 30 to 88 trials). Analysis of the data indicated that subjects engaged in a variety of behaviors as they attempted to determine whether the beam would balance. Not all common behaviors were observed in all subjects, suggesting that there may be no invariant sequence of behaviors that subjects must engage in while inducing the product-moment rule. There were three major findings observed in the interview data: (a) Most of the subjects seemed to develop and use the ratio rule before using the product-moment rule, (b) many subjects gave evidence of not encoding distance numerically during the first part of the session, and (c) most subjects seemed to employ specific information about previously experienced configurations in making decisions about balance problems.

Use of Rules of Limited Generality

At least 15 of the 22 subjects employed a quantitative rule that involved both weight and distance but was of limited generality before generating the product-moment rule. One subject generated and tested the hypothesis that balance would occur if the *sum* of weight and distance on each side of the fulcrum (e.g., a stack of three blocks one unit of distance away from the fulcrum would yield a sum of 4) was equal. Fourteen subjects explicitly verbalized a form of the ratio rule at some point during the interview (e.g., made a statement such as "It should balance because the stack on the left has twice as many blocks, but the stack on the right is twice as far from the center"). On the average, for these 14 subjects, the rule was first verbalized on Trial 21.1 (SD = 11.2), well before the criterion for the product-moment rule was reached. There was little difference in the number of trials to criterion for the 14 subjects who verbalized the ratio rule (M = 50.1, SD = 14.0) and the 8 who did not (M = 47.1, SD = 17.6).

In general, subjects who stated some form of the ratio rule did not do so on the first simple balance problem (i.e., a problem in which there were single, different-size stacks of blocks on each side of the fulcrum and the beam balanced) they encountered. The first simple balance problems were presented on Trial 7 (0003/0010) and Trial 11 (0004/0200). Only one subject stated the ratio rule on Trial 7, and only three others did so on Trial 11.

There seemed to be a tendency for subjects to state the ratio rule first on a trial for which the ratio was 2:1, suggesting that the rule was first generated in simple situations and then generalized. Ten of the 14 subjects who stated the ratio rule did so first on a trial for which the configuration was 0004/0200 or 0002/0100, despite the fact that all 10 had encountered one 3:1 ratio balance problem earlier (Trial 7) and six had previously encountered at least two simple balance problems in which the ratio was not 2:1.

Stating a form of the ratio rule correctly did not invariably result in correct predictions for all subsequently presented problems that could have been easily handled by use of the ratio rule. Apparently, subjects either did not initially learn the general form of the rule, did not employ the ratio rule as their exclusive heuristic for two-stack conflict problems, or did not always encode distance accurately enough to predict correctly. On the average, the 14 subjects missed 1.6 of the 6.6 simple balance problems they received between first verbalizing a form of the ratio rule and reaching criterion. In addition, they received an average of 2.4 simple imbalance problems (two-stack problems in which there was more weight on one side of the fulcrum and greater distance on the other but for which the ratios of weight and distance were not equal) and made an average of .8 errors on them.

Some error data suggest that subjects may have first learned the rule for one ratio and only later generalized it to others. Seven of the 10 subjects who first stated the ratio rule on a simple balance problem with ratio 2:1 subsequently made errors on other simple balance problems, but only one of them did so on a problem with ratio 2:1. Collectively, these subjects made 16 errors on simple balance problems after first stating the ratio rule, including 5 errors on problems with a 3:1 ratio, 5 on problems with a 3:2 ratio, and 4 on problems with a 4:1 ratio. The four subjects who first stated that ratio rule on a 3:1 or 4:1 ratio trial later committed 6 errors on simple balance trials, only one of which was on the same kind of trial.

Other data suggest that subjects may have made some errors on ratio trials by not judging distance carefully enough. As will be discussed later, subjects frequently did not use the markings on the scale to help encode distance numerically, and at least early in the session, tended to rely on ordinal or rough perceptual judgments. The hypothesis that errors on simple balance and imbalance problems were in part caused by the failure to encode distance numerically is supported by the fact that over all 22 subjects, only 4 errors on these types of problems were committed after subjects had given some indication that they were encoding distance numerically. Other suggestive evidence is provided by the fact that of the 11 errors made on simple imbalance trials after first verbalization of the ratio rule, 6 were committed on the configuration 0100/0200, whereas the others were committed on 2000/0030, 0003/0100, 1000/3000, 0003/0200, and 2000/4000. With the exception of the last configuration, each of the others could be made to balance by moving one of the stacks by no more than one-half unit. It is difficult to rule out the possibility, therefore, that crude perceptual judgments of distance were a contributing factor to this type of error.

In summary, more than two thirds of the subjects explicitly stated a quantitative rule involving both weight and distance before learning the product-moment rule. There is some evidence that subjects who verbalized a form of the ratio rule first tended to learn the rule for one ratio (usually 2:1) and then generalized it to others. Errors on simple ratio problems encountered after first verbalization of the ratio rule possibly occurred because (a) the rule had not yet become generalized, (b) subjects made crude judgments about distance, and (c) subjects also used other heuristics such as instance-based reasoning (to be discussed later).

Encoding of Distance

There was a substantial amount of evidence suggesting that subjects did not encode distance numerically early in the session. This evidence consisted of several types: (a) positive indication of beginning to use the scale to encode distance numerically, suggesting that the scale had not previously been used in this fashion, and (b) behavior or verbal statements indicating that subjects were encoding distance on the basis of crude perceptual estimates. In addition, some comments suggested that subjects encoded distance on an ordinal scale.

Scale use. Seventeen of the 22 subjects spontaneously gave evidence of counting or pointing to the unit marks on the scale in order to encode distance numerically. The mean trial on which they did so was 28.3 (SD = 10.6, range of Trial 14 to Trial 44). The remaining 5 subjects did not give clear evidence of using the scale until they received a fairly explicit prompt from the interviewer after Trial 40 (see Method section).

Thus, most subjects did not overtly use the scale until after Trial 20. This is to be contrasted with simply mentioning the use of distance either while thinking aloud or in explaining why a particular decision had been made. On the average, distance was first mentioned on Trial 6.2 (SD = 6.2, range of Trial 1 to Trial 24). However, the distribution was quite skewed. Thirteen of the 22 subjects mentioned distance within the first three trials, and only 2 took more than 15 trials to do so. The large average lag of 22 trials between commenting on distance and overtly counting distance units strongly suggests that subjects did not initially encode distance numerically.

Not all of the 17 subjects who spontaneously used the scale to measure distance seemed, at least initially, to realize the importance of doing so. Although 10 subjects continued to use the scale consistently after the first trial on which they gave evidence of counting, the remaining 7 did not, requiring an average of an additional 13.4 trials (SD = 8.0) before consistently using the scale to encode distance.

Of course, using overt counting of scale units provides an upper bound for the trial on which the subject begins to encode distance numerically, since it is possible to encode distance numerically without providing any evidence of doing so. Accordingly, explicit signs that subjects were not using the scale were sought.

Evidence of not using the scale. Nine of the 22 subjects gave positive evidence that they were not using the scale at some point between the first trial on which they mentioned distance and the first trial on which they overtly gave an indication of counting. Three of these subjects made comments that indicated they had just started to consider using the lines on the scale: One asked whether the lines were important (Trial 13), whereas the other 2 asked whether the distances between the marks were equal (Trials 33 and 36). Three different subjects indicated that they were judging distance in a crude perceptual manner, one commenting that she needed a different perspective and moving her head (Trial 28), and 2 others physically moving their chairs back and changing their lines of view (Trials 16 and 32). The remaining 3 subjects commented on the difficulty of judging distance, indicating that they did not realize that the scale was present to aid in those judgments (Trials 12, 25, and 34).

The fact that more than one third of the subjects gave explicit indications that they were not using the scale to measure distance, in combination with the relatively large number of trials it took to provide positive evidence of using the scale, strongly suggests that subjects did not use the scale early in the session. The fact that subjects mentioned distance early indicates that distance was encoded, but specifically how this was done is open to speculation.

Alternative hypotheses. If distance was not encoded numerically using the scale provided, there remain several possibilities about how it might have been encoded: (a) Subjects may have tried to estimate distance quantitatively on the basis of perceptual judgments, or (b) subjects may have simply encoded distance on an ordinal scale, ordering distances rather than attaching numerical values to them.

Several subjects gave clear evidence of trying to estimate distance precisely without using the scale; for example, one subject said, in response to 0004/0001, "It just looks like this [referring to the weight on the right] is four times the distance of these." However, there were a large number of comments that suggested subjects were indeed encoding distance on an ordinal scale. Eighteen of the 22 subjects made comments of this type at some point during the interview prior to giving explicit evidence of counting. Typical examples of such comments were (in response to 2000/2200) "It might balance because they're in pretty far [on the right] to have too much of an effect to go down," or "This one is farther over and it's less weight, and this one is closer and it's more weight." Statements of this type referred only to order relations, contained no reference to numbers of distance units, and were clearly distinguishable from statements that were later made after counting had begun, such as (in response to 1001/0020) "There's one, two, three, four, five, six [subject counts half-units], it's on the sixth line. Six and six is twelve" [subject adds the distance for each of the weights on the right]. Comments about order relations were noticeably absent from these later statements.

All comments suggesting the use of ordinal logic were tabulated for each subject. This coding included all trials in which a statement of an ordinal relation such as "farther out," "closer in," or "closer to the end" was made. Two types of cases were then excluded: (a) cases in which numerical encoding of distance was unnecessary (i.e., when weight was the same on either side of the fulcrum or when the greater weight was on the same side of the fulcrum as the greater distance), and (b) cases in which a specific reference was made to an earlier problem. There were, on the average, 5.5 trials per subject in which statements suggesting the use of ordinal logic were made, after the two types of trials mentioned above were excluded. Eliminating the data of one subject who made ordinal comments on 28 trials gives a mean of 4.4 such trials per subject ($SD = 3.5$). These data suggest that early in the session, subjects commonly dealt with distance in an ordinal fashion.

Use of Information About Previous Problems

There were an average of 7.4 trials per subject ($SD = 3.6$, range 2 to 16) that could be clearly documented on which subjects made or justified their predictions on the basis of a comparison with a previous problem or class of problems. Subjects made comments that they were using specific information about previous problems on more than 15% of trials, despite the fact that they were explicitly told that their goal was to learn a general rule. As mentioned in the Method section, the previous problem referred to was usually one of the immediately preceding problems.

There were three basic types of instance-based reasoning observed: (a) reasoning applied to a problem that differed from a previous problem by a single transformation, such as the addition of a block or the movement of a stack one unit of distance; (b) reasoning about a problem that differed from a previous problem by more than one transformation; and (c) reasoning in which the current problem was compared to a known ratio configuration. Valid and invalid methods of reasoning within each of the three basic types will be discussed below.

Single transformation. Subjects verbalized the use of instance-based reasoning on an average of 3.7 problems, in which the current problem was compared to a simple (one stack of blocks on each side of the fulcrum) problem that differed from it on the basis of a single transformation. There were three subtypes of reasoning which were sensible heuristics in that they produced correct answers some of the time. However, only one was consistently valid. The valid form of reasoning (verbalized on an average of 1.1 trials per subject) was that if a single transformation was made to a configuration that balanced, the new configuration should no longer balance. For example, if 0010/2000 is seen to balance, then 0100/2000 must not balance since only a single transformation has been made. One form of invalid reasoning (verbalized on an average of 1.0 trials per subject) was that if the present problem is similar enough to a previous one (differing by only one transformation), the outcome should be the same. The second type of invalid reasoning (1.5 trials per subject) was that if a single compensatory transformation is made to a configuration that did not balance, the new configuration should balance. Although this argument leads to correct predictions in some cases, it certainly does not do so in general.

Multiple transformations. There were an average of 2.3 problems per subject in which the prediction was made by comparing the current problem with a previous simple problem that differed from it by more than one transformation. The valid forms of reasoning (average of 0.7 trial per subject) included (a) If two stacks are equally distant from the fulcrum, equal numbers of weights added to each stack should not change the outcome (e.g., 0030/0200 should tip to the left given that 0020/0100 did so), and (b) given two stacks of weights at different distances from the fulcrum, an inverse proportional change in the number of weights on either side will leave the outcome unchanged (i.e., application of the ratio rule as a transformation). The invalid forms (1.5 trials per subject) included (a) Equal numbers of weights added to (or subtracted from) stacks at different distances from the fulcrum leave the outcome unchanged (e.g., if 0020/4000 balances, then so should 0010/3000), and (b) the addition (or subtraction) of a block to (or from) a stack in conjunction with the movement of the stack on the opposite side out (or in) one unit of distance will leave the outcome unchanged (e.g., if 0020/4000 balances, then so should 0200/5000).

Comparisons with known ratios. In addition, there were several types of reasoning in which the current complex problem (a problem with more than one stack on at least one side of the fulcrum) was compared with a problem that was known to balance because of the ratio rule. Valid forms of such reasoning (0.7 trial per subject) included (a) stating that the current problem should not balance because it differed in a critical way from one that would be predicted to balance using the ratio rule (e.g., 1100/4000 should not balance because the ratio rule would predict that 1000/4000 should balance), and (b) stating that the current problem should balance because it is the sum of two individual ratios each of which balances (e.g., 1020/4200 should balance because 1000/0200 and 0020/4000 both balance). In addition, if the current configuration was "similar enough" to a ratio configuration for which the outcome was known, the outcome was sometimes judged to be the same (0.7 trial per subject).

DISCUSSION

The major goal of the present study was to characterize some of the changes that occur when subjects are presented with a series of balance problems and given the task of inducing a rule that would allow them to predict the outcomes. In particular, we were concerned with whether the development of learning could be adequately characterized by Siegler's (1976) rule-based hierarchy of models, or whether heuristics that do not fit neatly into this hierarchy, such as differential encoding of distance, use of the ratio rule, and reasoning on the basis of previous instances of problems, must also be considered. Because our data strongly suggest that such heuristics are indeed used by subjects, Siegler's (1976) hierarchy does not adequately describe the variety and complexity of reasoning processes that subjects engage in while attempting to induce a general rule.

A Well-Defined Sequence?

It is unlikely that any simple stage analysis can characterize the changes in knowledge states in more than a superficial manner. Our analysis of the protocols does not depict the subject as relentlessly progressing through a well-defined sequence of levels until the product-moment rule is reached. Distance may be encoded quantitatively on one trial but not on the next. The ratio rule may not be applied at all or may be applied to make a correct prediction on one problem and not be applied when a similar problem is presented a few trials later. The strategy employed depends on the particular problem and on problems that were encountered earlier.

Model-Determined Encoding of Distance

The protocols suggest that initially, subjects often rely on a relatively primitive encoding of distance. After they have made predictions about a number of simple balance problems, they tend to progress to the use of some mixture of the ratio rule (where appropriate) and instance-based reasoning and begin to consider more complex hypotheses. It seems likely that the level of encoding employed by subjects is largely determined by the heuristics or models they are using. The argument for this is twofold. First, our subjects were of college age, and it is extremely likely that they thought of distance as a quantitative concept in general. Thus, any failure to encode distance in a quantitative fashion more plausibly represents a failure to apply an existing concept of quantitative distance rather than the lack of the concept. Second, there were occasions in which subjects used qualitative encoding after having earlier given some indication of encoding distance quantitatively, such as counting scale units or using the ratio rule. It seems plausible that some occurrences of qualitative encoding resulted from the use of more qualitative models, such as those involving instance-based reasoning. The pattern of reasoning seems similar to that engaged in by a mathematician or logician when he or she attempts to understand a problem on an intuitive, informal level before trying to formalize a hypothesis.

The hypothesis that the distance encoding employed by subjects will depend on the heuristics or models they use contrasts with conclusions that might be drawn from Siegler's (1976) work on encoding with 5- and 8-year-olds. Siegler demonstrated that training children to encode both weight and distance significantly improved their performance on balance beam problems. Although quantitative encoding is necessary for an adequate understanding of the balance beam, it is not sufficient for understanding. The explanation of progress in understanding the balance beam in terms of learning how to encode distance quantitatively has less appeal when applied to older subjects who already know how to encode distance quantitatively.

Instance-Based Reasoning

Making a prediction about a problem by comparing it to previously experienced problems seems to be outside the scope of a rule-based model of learning. In fact, one might think that when subjects were instructed to generate a rule, the demand characteristics of the task would preclude reasoning based on individual instances. However, the prevalence of such reasoning is less surprising within the context of a growing literature concerned with the learning of complex concepts. Studies have demonstrated that, in learning to classify instances of a concept by a complex rule or set of rules, for example, the pronunciation of English words (Brooks, 1978) or the learning of syntactic structures (Reber, 1976) is most parsimoniously accounted for by a theory that claims subjects categorize new instances by comparing them to specific old instances. The product-moment rule may be sufficiently complex that basing a prediction on previous instances involves less strain on cognitive resources than immediately trying to infer a quantitative law.

If one assumes that instance-based reasoning was prevalent, it is possible to develop an explanation for why subjects were not more consistent in their use of the ratio rule once they had first verbalized a form of the rule. On a given problem, subjects may have used whatever information was available and seemed appropriate for that problem. In addition, there was some evidence that subjects first learned the ratio rule for a specific ratio and later generalized it. When subjects first verbalized a form of the ratio rule, they rarely made comments suggesting that the rule should hold for all ratios, and in some cases expressed doubts that the relationship observed in a few instances was generalizable. For example, one subject said in considering whether the ratio rule verbalized earlier for a smaller ratio would hold for 4:1, "I know we kept the proportion the same, but I thought that there was a point at which you went one too many down here [referring to distance] and the ratio didn't stay the same just because it was so far out on the end. You want to put one more block on there and move that one more? [requesting that the interviewer modify the problem from 1000/4000 to 10000/5000]. Oh, you mean it's a constant rule, it doesn't change?" [after observing that the beam still balanced].

Why Do Subjects Progress?

The documentation and description of a change in understanding does not in itself provide an explanation of the mechanism producing the change. At most, the protocol provides a record of the stopping places on the journey. However, we believe that we will have a better chance of ultimately understanding how subjects induce the general rule if we can better specify what they do during the transitional period by documenting the changes in their understanding and inferring the kinds of heuristics they employ. At the very least, it is clear that our training procedure was sufficient, in that all of our subjects were able to generate the product-moment rule within a reasonable amount of time.

This finding contrasts with the results Siegler (1976) obtained with younger subjects that indicated relatively few subjects induced the product-moment rule. Although it is possible that the difference in results was due to the age of the subjects, there are several procedural differences between these training studies that seem relevant. First, subjects in the present study took 49.0 trials to reach criterion, whereas subjects in Siegler's observation condition received only 36 trials. Thus, Siegler's observation condition may not have provided enough trials to be effective. The second, and more interesting possibility is that a critical difference may have been that in the present study subjects were asked to generate predictions for each problem and to justify these predictions if possible, while Siegler's subjects merely observed the problems and the outcomes. More active involvement with the problems may lead to more active hypothesis formation and hence faster learning. This interpretation is supported by recent research by Lewis and Anderson (1985), who found in a study of the acquisition of problem-solving operators that subjects learned correlations between problem features and operators only when they were forced to make and test explicit hypotheses. Clearly, active learning should be an important topic of research in the future. Siegler and Klahr (1982) also conducted a training study in which the majority of college-age subjects were able to learn to provide predictions consistent with Model 4 after being presented with a sequence of balance-scale feedback problems, provided they received either external memory aids (a sheet of paper with schematic representations of each problem and its outcome), quantified encoding (a procedure in which the mathematical nature of the task was highlighted), or both. Siegler (personal communication, April, 1985) has suggested that the effectiveness of external memory aids may be explained in terms of making specific instances more available.

Educational Value of the Concept of Balancing

The topic of how people come to understand the balance beam gains significance when one considers that the balancing schema might influence the acquisition of mathematical concepts. One recent study has indicated that knowledge about the balance beam may facilitate understanding of the superficially unrelated concept of the mean (Hardiman, Well, & Pollatsek, 1984). Nearly all college students know the standard algorithm for computing the mean of a set of scores (i.e., add the scores and divide by the number of scores) and can apply it correctly when the scores provided are based on equal numbers of observations. However, relatively few adults can successfully compute the overall mean when given two subgroup means based on different numbers of scores. They do not "weight" the means in proportion to the number of scores in each group, but rather treat the group means as though they were based on equal numbers of scores (Pollatsek, Lima, & Well, 1981).

Pollatsek et al. (1981) suggested that subjects who failed to solve weighted mean problems lacked "analog" knowledge of the mean. Such analog knowledge might involve a visual or kinesthetic image of the mean as a balance point, a commonly used metaphor in textbooks. Given two subgroup means based on different numbers of observations, the weighted mean or "balance point" must lie closer to the subgroup mean based on the larger number of observations because that mean "weighs" more. Hardiman et al. (1984) assessed college students' knowledge of the concept of balancing and of the mean. Subjects who were consistently able to predict the outcomes of balance problems correctly were also able to solve weighted mean problems correctly and could represent weighted mean problems on the balance beam. More important, subjects who initially performed poorly on both balance beam and weighted mean problems and who were given training on balance problems (but not on problems specifically having to do with the mean) subsequently were able to solve weighted mean problems better and with greater understanding than control subjects who had not received this training. Thus, the understanding of an important concept such as the mean can be enhanced through training with balance problems. An interesting question is whether acquisition of the product-moment rule is necessary in order for transfer to take place.

Conclusions

Our study of how subjects attempt to make correct predictions about balance problems leads us to form several conclusions. Reasoning before the product-moment rule has been learned is complex, involving heuristics such as ordinal encoding of distance, reasoning from previous problems, and using quantitative rules of limited generality. Not all heuristics seem to be employed by all subjects, indicating that there may be different paths toward acquiring the product-moment rule. Thus, it is not likely that a simple stage analysis or hierarchy of rules can adequately reflect the dynamics and complexity of intermediate stages of learning. The use of both limited rules like the ratio rule and instance-based reasoning does not fit into the models of Siegler's (1976) rule-based hierarchy, nor does it conform to methods of discovery employed by some artificial intelligence programs; for example, a program receives data in a tabular form and systematically reduces it to a single empirical law (Langley, 1981).

Brooks's (1978) work suggests that in relatively complex situations, categorization based on comparisons with previously experienced exemplars may represent less of a load on cognitive resources than an attempt to abstract a rule. The central role of critical examples has been acknowledged within the domain of mathematics (Michener, 1978) and in the context of learning to classify instances generated by a complex rule or set of rules (Brooks, 1978; Reber, 1976), but not within the context of inducing a relatively simple physical law. The documentation of the use of instance-based reasoning in the present study suggests that reasoning from salient examples may be a rather general heuristic employed in many types of reasoning, including judgments about frequency and probability (e.g., Kahneman, Slovic, & Tversky, 1982), understanding mathematics (Michener, 1978) and physics (Clement, 1981), and deducing syntactic structure (Reber, 1976), as well as in the discovery of physical laws. A major question facing the study of problem solving is exactly how the conscious discovery of rules emerges from such instance-based reasoning.

ACKNOWLEDGMENTS

This research was supported by funds from National Science Foundation Grants SED-8113323 and SED-8016567.

We thank Charlene Rheaume, Susan Lima, and Gale Sinatra for their assistance in various phases of protocol transcription and analysis. Jere Confrey provided helpful suggestions in the early stage of data analysis. We are also indebted to Carolyn Mervis for her comments on the paper.

REFERENCES

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 170–216). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Clement, J. (1981). *Analog generation in scientific problem solving.* Paper presented at the 3rd annual meeting of the Cognitive Science Society, Berkeley, CA.

Hardiman, P. T., Well, A. D., & Pollatsek, A. (1984). The usefulness of a balance model in understanding the mean. *Journal of Educational Psychology, 76,* 792–801.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic.

Jackson, S. (1965). The growth of logical thinking in normal and subnormal children. *British Journal of Educational Psychology, 35,* 255–258.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

Klahr, D., & Siegler, R. S. (1978). The representation of children's knowledge. In H. W. Reese & L. P. Lipsett (Eds.), *Advances in child development and behavior* (Vol. 12, pp. 62–116). New York: Academic.

Langley, P. (1981). Data-driven discovery of physical laws. *Cognitive Science, 5,* 31–54.

Lewis, M. W., & Anderson, J. R. (1985). Discrimination of operator schemata in problem solving: Learning from examples. *Cognitive Psychology, 17,* 26–65.

Lovell, K. (1961). A follow-up study of Inhelder and Piaget's "The growth of logical thinking." *British Journal of Psychology, 52,* 143–153.

Michener, E. R. (1978). Understanding understanding mathematics. *Cognitive Science, 2,* 361-383.

Pollatsek, A., Lima, S. D., & Well, A. D. (1981). Concept or computation: Students' understanding of the mean. *Educational Studies in Mathematics, 12,* 191-204.

Reber, A. S. (1976). Implicit learning of synthetic languages: The role of instructional set. *Journal of Experimental Psychology: Human Memory and Learning, 2,* 88-94.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8,* 481-520.

Siegler, R. S. (1978). The origins of scientific reasoning. In R. S. Siegler (Ed.), *Children's thinking: What develops?* (pp. 109-151). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Siegler, R. S., & Klahr, D. (1982). Why do children learn? The relationship between existing knowledge and the acquisition of new knowledge. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 2, pp. 121-211). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Strauss, S., & Levin, I. (1981). [Commentary on and reply to *Developmental sequences within and between concepts*]. *Monographs of the Society for Research in Child Development, 46* (2, Serial No. 189).

Wilkening, F., & Anderson, N. H. (1982). Comparison of two rule-assessment methodologies for studying cognitive development and knowledge structure. *Psychological Bulletin, 92,* 215-237.