

Siegler, R. S., & Robinson, M. (1982). The development of numerical understandings. In H. W. Reese & L. P. Lipsitt (Eds.), *Advances in child development and behavior: Vol. 16* (pp. 242-312). New York: Academic Press. (PDF #1 of 2.)

THE DEVELOPMENT OF NUMERICAL UNDERSTANDINGS¹

Robert S. Siegler and Mitchell Robinson

DEPARTMENT OF PSYCHOLOGY
 CARNEGIE-MELLON UNIVERSITY
 PITTSBURGH, PENNSYLVANIA

- I. INTRODUCTION 242
- II. AN INITIAL STUDY OF NUMBER CONSERVATION 244
- III. PRESCHOOLERS' KNOWLEDGE OF COUNTING 250
 - A. COUNTING FROM ONE 250
 - B. COUNTING ON FROM A POINT BEYOND ONE 256
 - C. MODELS OF THREE LEVELS OF COUNTING EXPERTISE 259
- IV. PRESCHOOLERS' KNOWLEDGE OF NUMERICAL MAGNITUDES 267
 - A. EXISTING RESEARCH ON ADULTS AND CHILDREN 267
 - B. PRESCHOOLERS' NUMERICAL MAGNITUDE COMPARISONS 271
 - C. VERBAL LABELING OF NUMBERS 275
 - D. EFFECTS OF TEACHING A LABELING STRATEGY 278
 - E. THREE MODELS OF NUMERICAL MAGNITUDE COMPARISON 280
- V. PRESCHOOLERS' KNOWLEDGE OF ADDITION 287
 - A. EXISTING RESEARCH ON CHILDREN 287
 - B. PRESCHOOLERS' ADDITION STRATEGIES 290
 - C. A MODEL OF STRATEGY CHOICE IN ADDITION 296
- VI. CONCLUSIONS: THE DEVELOPMENT OF NUMERICAL KNOWLEDGE 299

REFERENCES 308

¹This research was supported in part by NICHD Grant HD-15285 and by a grant from the Spencer Foundation. Thanks are due to Dr. Taylor and the teachers of the Carnegie-Mellon Children's School, and to Mrs. Cohen and the teachers of the Carnegie-Mellon Day Care Center. Also deserving considerable gratitude are Cindy Zaks who was the experimenter in all of the studies and Vickie DeRose and Amy Laird who typed more than a reasonable number of drafts of the manuscript. Finally, a number of colleagues read and commented on earlier versions of the article: Diane Mierkiewicz, Janellen Huttenlocher, Hayne Reese, and Roman Taraban. Their help is gratefully acknowledged.

Courtney, 1969), most children can succeed on variants of it by age 2 or 3 (Bever, Mehler, & Epstein, 1968; Bryant, 1974; Gelman, 1972). To cite another illustration, Trabasso (1978) listed some of the task variables that have been found to influence class inclusion performance: typicality of subordinate class exemplars, presence of contrasting subordinate classes along with the two classes referred to in the question, ratio of majority to minority subclass, whether the members of the sets are physically present, and whether the subordinate classes are quantified before or after pictures are shown. The large effects of many of these variables make it extremely difficult to decide which form of the task is optimally suited for assessing understanding (Flavell, 1971; Siegler, 1981a).

The situation becomes even more complicated when we consider not just variants of a single task but also the many possible tasks that might reasonably be said to correspond to any concept. Rather than assessing knowledge of numbers by the number conservation problem, we might assess it in terms of ability to count objects, to compare numerical magnitudes, to understand the relationship between arithmetic and algebra, to understand number theory, and so on. Again, there is no simple principled way to choose.

The issue can be considered at a very general level. Braine (1959) and Brown (1976) have made quite eloquent statements advocating specific criteria for defining conceptual understanding. Braine argued for a criterion of initial competence, Brown for a criterion of stable usage. The dilemmas that each of these proposals lead to suggest that no single standard of conceptual understanding can be adequate. Consider Braine's (1959) statement:

It is clear that if one seeks to state an age at which a particular type of response develops, the only age that is not completely arbitrary is the earliest age at which this type of response can be elicited using the simplest experimental procedure. (p. 16)

This statement is entirely reasonable, as far as it goes. When one considers the long time period separating initial and mature understanding of concepts, however, a paradox becomes evident. Adopting the initial competence criterion puts us in the position of saying that many concepts develop at relatively young ages, yet of also saying that children fail many reasonable indices of understanding for years thereafter. Stated another way, much, perhaps most, of conceptual growth would be seen as occurring after the concept "develops."

Brown (1976) implicitly suggested an alternative criterion in her discussion of the development of seriation:

Under optimal circumstances, they can indeed seriate a succession of pictures representing a time course. . . . Yet how robust is their concept of succession? Is it truly operative according to the defining features argument? The answer must be "no," for their concept of order appears to be extremely fragile and is disrupted by seemingly trivial changes in the optimal task. (p. 77)

I. Introduction

The purpose of this article is to explore ways in which we can characterize people's understandings of concepts. In particular, we shall attempt to characterize very young children's understandings of numbers by examining a variety of their numerical skills, by inferring representations and processes that might give rise to each of these skills, and by integrating the models arising from each task domain to build a general model of preschoolers' knowledge of numbers.

At one time, not very long ago, there was a large degree of consensus as to how best to describe children's knowledge of concepts. Piaget had identified tasks that came to be accepted as indices both of children's general cognitive levels and of their understanding of specific concepts. Children were said to understand the concept of classes when and only when they could succeed on the class inclusion task, to understand the concept of ordering when and only when they could succeed on the seriation task, to understand the concept of perspective when and only when they could succeed on the three-mountain task, and so on. Even though Piaget himself examined numerous tasks before reaching his conclusions, many other investigations relied on a single one of his tasks as their sole index of understanding of each concept.

Accompanying the view that understanding could be assessed by performance on a single task was the view that children possessed a single understanding to measure. This view was reflected in the multitude of studies in which researchers attempted to establish *the* age at which children master particular concepts and *the* order in which children master different concepts. It was also reflected in the titles that investigators, both Piagetians and non-Piagetians, chose for their reports. Illustratively, two of the most influential monographs on children's understanding of number concepts have been those of Piaget (1952) and Brainerd (1979). Piaget titled his book *The Child's Conception of Number* Brainerd titled his book *The Origins of the Number Concept*. The use of the singular in the terms "conception" and "number" and of the definite article together with the singular in the phrase "the number concept" is striking. The implication is that there exists a single number concept to be understood and that children have a particular concept or conception of it.

Mathematicians and philosophers have long debated whether there exists any single concept of number to be understood, but the results of the past 20 years of research in developmental psychology have rendered completely untenable the position that children possess a single understanding of number or of other complex concepts. Even within a single basic problem, wide variability in performance, depending on the details of the task, has been the rule. To cite one illustration, although most children do not succeed on Piaget's number conservation problem until age 6 or 7 years (Beilin, 1968; Miller, 1976; Rothenberg &

The paradox that is implicit in Brown's observation is pointed out in Braine's comment. What exactly does a child understand when he or she can use a concept in some situations but fails to qualify as having an operative understanding of it? As Braine suggested, it does seem arbitrary to identify understanding with anything other than the earliest form of understanding; however, it seems misleading to identify understanding with the earliest form of understanding.

How then can we study conceptual development? The above critique suggests a three-stage procedure. First, one examines performance on a variety of tasks corresponding to different aspects of a concept. For the same reasons that it seems desirable for psychometricians to sample numerous content domains in order to infer children's status on aggregate-level constructs such as intelligence, it seems desirable for cognitive and educational psychologists to sample content within any given concept to infer children's conceptual understanding. Second, one characterizes the representations and processes that people use to perform each particular task. The reasons for formulating models that distinguish between representations and processes have been discussed previously by Anderson (1976, 1978) and will be discussed further in a later section of this contribution. Third, one integrates the findings from each task into a general characterization of knowledge sufficient to derive each of the particular representations and the processes applied to it. Information-processing models have been criticized for being overly task specific, for not addressing the connections between how people perform any one task and other aspects of their knowledge (Neisser, 1976; Strauss & Levin, 1981). Producing a model that revealed the relationships among several tasks within a given conceptual domain would constitute at least a first step toward meeting this criticism. In the remainder of this article, we shall describe our efforts to apply this research strategy to analyzing preschoolers' concepts of numbers.

II. An Initial Study of Number Conservation

A study of Piaget's number conservation problem (Siegler, 1981a, Experiments 3 and 4) was instrumental in convincing us of the need to adhere to the above-described procedures in order to characterize children's knowledge. This study began with a consideration of the role of transformations in conservation problems. Analysis suggested that transformations play a crucial role within the conservation concept. The one sure way to determine whether the value of a particular quantitative dimension will be preserved in a situation is to know the type of transformation that will be performed. Adding something to that dimension necessarily results in more, subtracting something results in less, and neither adding nor subtracting anything results in the same amount as before.

Piaget and subsequent conservation researchers focused almost exclusively on

transformations that do not affect quantity: pouring water, molding clay, moving objects apart, and so on. It would seem, however, that transformations that do affect quantity—addition and subtraction—are of at least equal importance. A conserver might reasonably be expected to understand not only that spreading out a row of objects leaves the number of objects unchanged, but also that spreading them and adding an object means that the row now has more objects than before and that spreading them and taking away an object means that it now has less.

In Siegler's (1981a) study, 3- to 9-year-olds were presented number conservation problems in which the starting configuration always had two equally numerous and equally lengthy rows of objects. Transformations varied in their effects on numerosity and length. Some problems involved adding objects to a row, some problems involved subtracting objects from a row, and some involved neither adding nor subtracting objects. Some problems involved lengthening the transformed row of objects, some involved shortening the row, and some involved moving the objects but ultimately returning the row to its original length. Finally, on some problems the rows had few objects (2-4), and on others they had many (7-9). In all, 48 problems were presented. The pattern of correct answers and errors on these problems was sufficient to allow us to induce the rules that children were using to perform the task (see Siegler, 1976, 1978, 1981a, for descriptions of the rule assessment approach that was used).

Most children used one of five rules in approaching the conservation problems. These rules incorporated progressively more complete subordinations of perceptual to transformational criteria. As shown in Fig. 1A, the youngest children, most of them 3-year-olds, chose the longer row as having more objects in almost all cases—whenever the problem involved a large number of objects, and even with small arrays, when nothing had been added or subtracted. When the number of items was small and something had been added or subtracted, however, they relied on the type of transformation; addition meant that there was more and subtraction that there was less, regardless of the relative lengths of the rows.²

Somewhat older children used a simplified version of this approach (Fig. 1B). When only a few items were presented, they always decided on the basis of the type of transformation. However, when many objects were presented, the children chose the longer row as having more.

Yet older children extended the transformational approach one step further (Fig. 1C). Now, either if the rows had few items, regardless of the transformation, or if they had many and something had been added to or subtracted from

²Children using Rules I, II, and III may have used set size and type of transformation cues as indices of when to count rather than as direct bases of judgment. Within this interpretation, the entire Rule IV procedure would replace the bottom-most diamond and branches on the left side of Rules I, II, and III. Available evidence did not allow discrimination between these two interpretations.

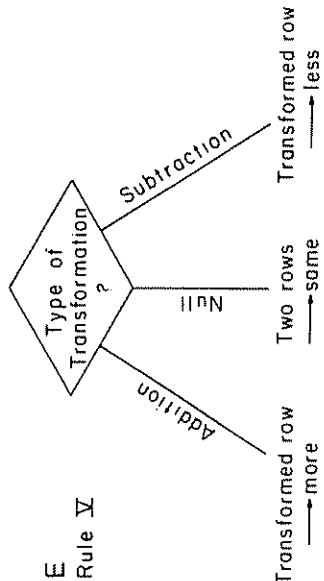
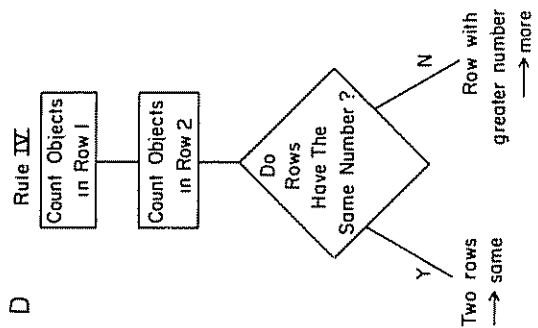
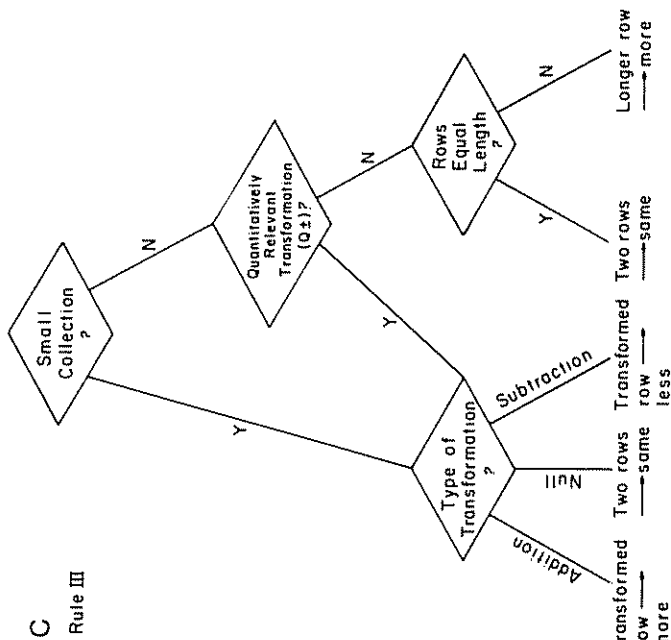
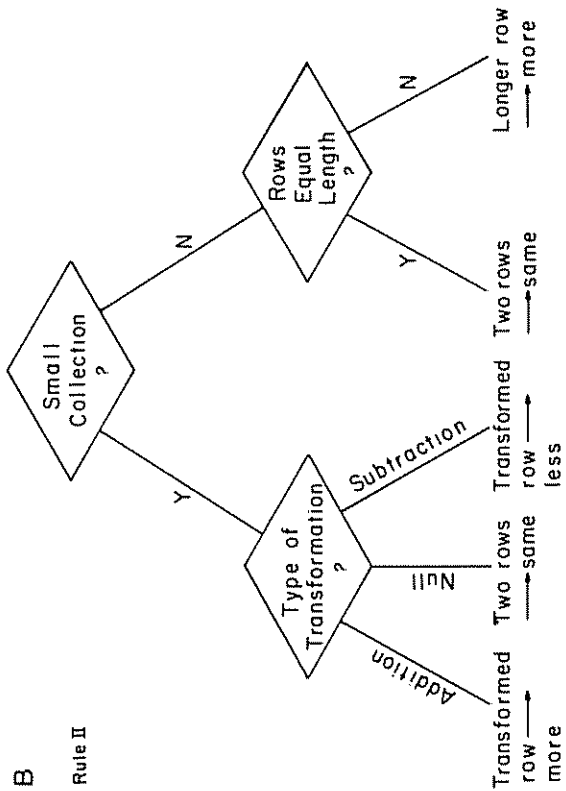
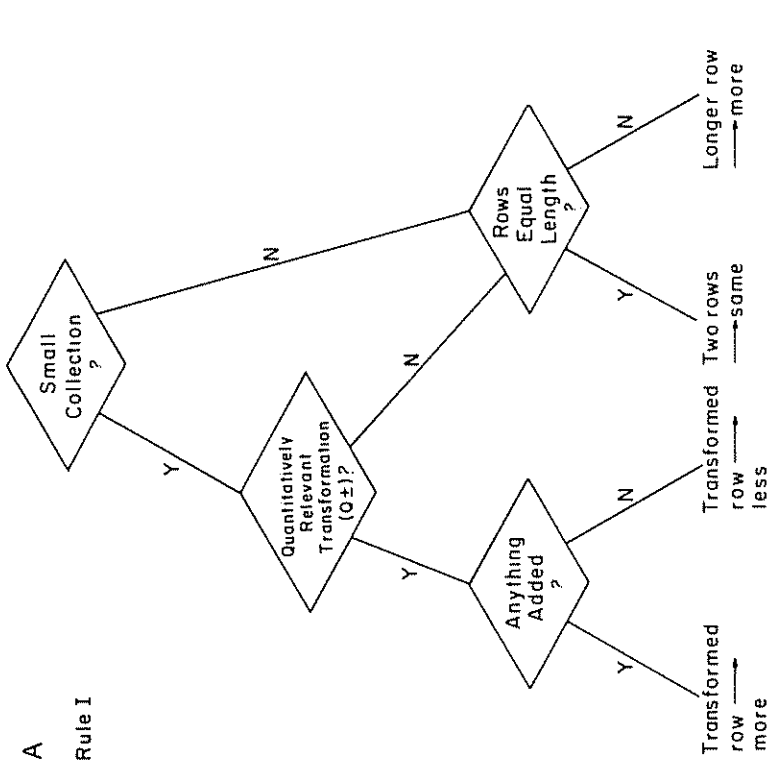


Fig. 1. (C) Rule III; (D) Rule IV; (E) Rule V

Fig. 1. Rules on number conservation task: (A) Rule I; (B) Rule II.

one of them, the type of transformation was the determinant of the conservation judgment. If the rows had many items and nothing had been added or subtracted, however, children still chose the longer row as having more. Interestingly, this last holdout from the transformational criterion is precisely the traditional conservation of number problem.

Still older children came to use Rule IV on both small and large sets; they answered all types of number conservation problems correctly. This achievement did not conclude the development of the concept, however. Rule IV children solved number conservation problems by counting or pairing the objects in the two rows (Fig. 1D); Rule V children realized even without counting or pairing that adding something necessarily meant that that row had more, that subtracting something necessarily meant that it had less, and that doing neither meant that it necessarily had the same number of objects as before (Fig. 1E).

The same 3- to 9-year-olds whose understanding of number conservation was assessed were also tested on the liquid and solid quantity conservation tasks. Two findings suggested that understanding of number conservation was crucial to the development of understanding of the other two conservation problems. First, no children consistently solved liquid and solid quantity conservation problems who did not also consistently solve number conservation problems. Second, among children who consistently solved number conservation problems, those who justified their responses on those problems in terms of the type of transformation were much more likely to solve the other two types of conservation problems than children who justified their responses on the number problems in terms of counting or pairing of objects in the two rows. These findings suggested an overall model of conservation acquisition. After progressing through rules that will solve some but not all number conservation problems, children come to be able to solve all problems by using the external referents of counting and/or pairing. Later, they note that these tests always indicate that when something has been added, there are more objects than before, that when something has been subtracted, there are fewer, and that when nothing has been added or subtracted, there are the same number. Therefore, they come to rely on the type of transformation to solve the number conservation problems. Finally, the children apply what they have learned about transformations in the number context to other domains involving transformations but not offering simple external referents, specifically conservation of liquid and solid quantity.

This experiment persuaded us of several points. First, it illustrated just how difficult defining conceptual understanding is. If understanding of number conservation were defined as "in at least some situations, relying on the type of transformation despite opposing perceptual cues," then Rule I would be the point of understanding. If understanding were defined as "for the traditional number conservation transformation, in at least some situations relying on the type of transformation despite opposing perceptual cues," then Rule II would be chosen. If it were defined as "ability to consistently solve number conservation prob-

lems," then Rule IV would be chosen. If it were defined as "an abstract understanding of the quantitative effects of transformations that does not depend upon requantification and that can be transferred to other types of conservation problems," then Rule V would be chosen. Each one of these definitions seems to us sensible within particular theoretical contexts; no one of them seems absolutely more sensible than the others.

A second lesson that we derived from the conservation study was the desirability of separating children's representations of information from the processes that they apply to their representations. The rule models specified the processes that children used to solve the conservation problems but only implicitly addressed the content upon which the processes operated. How many objects constituted a small number and how many a large one was never spelled out. Neither was the form of the knowledge about transformations that children transferred from the number conservation context to the liquid and solid quantity ones. For some issues, such as how many objects constitute a small number, specifying the additional information would have been awkward but possible. For other issues, such as how knowledge of transformations is represented, it is unclear how the information could have been included within the rule format.

Anderson's (1976) suggestion that representations and processes (declarative and procedural knowledge) be separately specified held out the promise of more precise yet also more economical description. Within his framework, diverse behaviors might be characterized as the products of different processes being applied to a single strategy-free representation. Specifying the representation would force us to consider the structure of the relevant knowledge domains and also the interrelationships among the knowledge domains drawn upon in solving different tasks. Specifying the processes would force us to consider the ways in which this information is manipulated to meet the demands of particular tasks. Therefore, in the series of studies presented here, rather than only concentrating on the processes (rules) that children use in solving numerical problems, we also attempted to specify their representations of numbers.

A third lesson that we drew from the conservation experiments is that even in a single problem such as number conservation, a variety of other types of skills might be crucial to children's success. Consider the skills that were likely involved in using the correct performance rules, Rules IV and V. Children using Rule IV were hypothesized to judge the results of the transformations by counting each row of objects. After counting, the children needed to compare the magnitudes of the numbers to determine which row, if either, had more. Similarly, children using Rule V needed to know about the directional effects of addition and subtraction in order to decide, without enumerating, which row had more.

The likely involvement of these skills in the mastery of number conservation was one incentive to single them out for study. Other motivations were even more compelling, though. Counting, comparing, and adding are basic mathemat-

ical skills.³ They are activities in which very young children are interested, in which they frequently engage, and at which they possess considerable skill; thus, they are useful for illustrating what young children *do* know (Brown, 1978; Gelman, 1978; Ginsburg, 1977). Each of the skills generates a large amount of observable behavior in preschoolers, though not necessarily in older children and adults; such visible manifestations of reasoning provide at least a translucent window for inferring the representations and processes that underlie behavior. The skills have numerous possible interrelationships, in addition to the common involvement in number conservation. For example, Groen and Parkman (1972) postulated that when people add two integers, they compare the numbers to determine the larger ones and then count up the number of times indicated by the smaller (Groen & Parkman, 1972). Finally, the three types of skills are sufficiently diverse that data on them, together with the already collected data on number conservation, should allow reasonably broadly based inferences about children's knowledge of numbers. In the next three sections of this article, we describe experiments intended to reveal how knowledge of counting, numerical magnitudes, and addition develops during the preschool period.

III. Preschoolers' Knowledge of Counting

A. COUNTING FROM ONE

Perhaps the first experience that most children have with most numbers is in the context of the counting string. Although words denoting small numbers may have many semantic referents (e.g., two ears, three people in our family, four years old), words denoting larger numbers such as 11, 15, or 23 almost certainly do not. Learning about numbers in the counting context may help children learn about them in other contexts as well. Illustratively, Pollio and Whitacre (1970) reported that the length of preschoolers' counting strings is an excellent predictor of their ability to establish 1-1 correspondence, to divide objects into equally numerous sets, to insert the missing number into a series, and to count on from an arbitrarily chosen point within the number string. The first goal of this series of experiments, therefore, was to establish the representation and process that children apply in abstract counting (i.e., use of the number string in the absence of objects).

To date, researchers have reported three major efforts to describe young children's abstract counting: Fuson and Richards (1982), Ginsburg (1977), and Greeno, Riley, and Gelman (1981). The alternatives posed by Ginsburg and by

³In the report presented here, the only research on arithmetic that will be discussed involves knowledge of addition. Research on subtraction is now being conducted and is certainly part of young children's knowledge of numbers.

Greeno and co-workers are quite explicit in their hypotheses about the underlying representation of the number string and therefore are of the greatest interest in the present context. Ginsburg (1977) postulated a great deal of structure in children's representations of the number string, beginning in the teens:

The beginning of the sequence—the first 12 numbers or so—is completely arbitrary. There is no rational basis for predicting what comes after a certain number. Therefore, children have to memorize the smaller numbers in rote fashion. After a period of time, they discover that the numbers after about 13 contain an underlying pattern. Using it, children develop a few simple rules by which to generate the numbers up to about 100. (p. 9)

In contrast, Greeno and co-workers postulated a representation without any particular structure among the numbers. They hypothesized that numbers are simply connected by the "next" relation. Thus, as they noted, their model resembles Peano's (1899) theory of numbers in its emphasis on the successor relationship.

The models of Ginsburg and Greeno and co-workers represent the logically possible extremes in the amount of structure that children might impose on the number string. A third, intermediate possibility also existed: that children might detect and use the relatively transparent structure that appears in the number string beyond the number 20 but not the less obvious structure that is present in the teens. These three models predicted counts that differed in a large number of qualities. Consider just one, stopping points. If, as in the model of Greeno and co-workers, only "next" connections bind the numbers within the number string, there would be no reason for children to be more likely to stop at any one point than at any other; thus, the distribution of stopping points should be relatively flat across the number string. Alternatively, if, as in Ginsburg's model, children treat the teens as the first decade with a repetitive structure, then they might be expected to stop differentially at points where the structure did not indicate the name of the next number: 19, 29, 39, and so on. Finally, if, as in the third possibility described above, children abstract the structure of the number string only beyond the number 20, then they would be expected to stop relatively often at 29, 39, and 49, but not at 19.

Neither Ginsburg nor Greeno and co-workers presented data, beyond anecdotal reports, that supported their model or failed to support other models. We could not locate any other detailed descriptions of young children's abstract counting either. Therefore, our first experiment was an effort to obtain a data base from which to generate one or more detailed models of the knowledge underlying preschoolers' abstract counting.

1. Method

The children who participated in this and in most of the subsequent experiments were 13 3-year-olds, 19 4-year-olds, and 10 5-year-olds. Among the 3-year-olds

of the departures: their prevalence (the percentage of children in each group who exhibited the departure at least once), their frequency (the number of departures relative to the total number of counts by children in the group), and their quality (the specific types of departures that were observed).

b. Omissions. In order to quantify the frequency of omissions, we treated each omission as a single instance, regardless of the span of omitted numbers (i.e., a skip from 14 to 16 was treated in the same way as a skip from 29 to 60). As Table I indicates, almost all children omitted at least one number from their counts. The frequency of omissions was found to decrease significantly, though not hugely, with the child's expertise in counting. The ratio of omissions to numbers counted was 5% for children who stopped before reaching 20, 3% for children who stopped between 20 and 99, and 1% for children who stopped beyond 99.

The contrasts among groups were much more striking in the quality of the omissions. None of the children in the least expert group skipped 10 or more consecutive numbers, but roughly half in the two more expert groups did. The ratio of decades skipped to decades counted was 0 in the group that stopped counting below 20, .12 in the group that stopped between 20 and 99, and .07 in the group that counted beyond 99. All but 3 of the 43 decade omissions involved jumps from a number ending in '9' to the beginning of another decade (e.g., 27, 28, 29, 50). This finding resembles previous descriptions of adults' learning to count in bases 2, 3, and 4 (Pollio & Reinhardt, 1970). There, as well as here, the majority of large omissions involved jumps from the highest value of the base unit to a too-advanced later point.

c. Repetitions. Almost all of the children sometimes repeated numbers (Table I). However, both the frequency and the quality of the repetitions changed markedly with expertise. Viewing each repetition as a separate instance, regardless of the length of the repeated list, we calculated for each counting group the ratio of repetitions to numbers counted. This ratio decreased from 26% for those whose counts did not reach 20 to 2% for those whose counts ended between 20 and 99 to 1% for those whose counts reached 100 or more.

The quality of the repetitions also varied markedly with expertise. Only 3% of the repetitions of children in the 1-19 group and 0% of the repetitions of children in the 100+ group involved more than three numbers. In contrast, 22% of the repetitions of those who stopped counting between 20 and 99 involved at least nine numbers. All but two of the extended repetitions involved a number ending in '9' as the stepping-off point; the child would reach the '9' number and then regress to a number ending in either '0' or '1' in an already completed decade.

d. Nonstandard numbers. A sizable minority of the children used at least one nonstandard number in their counts (Table I). The prevalence of use of these

TABLE I
Data on Counting

Measure ^a	Group		
	1-19 (N = 10)	20-99 (N = 26)	100+ (N = 6)
Stopping point			
Counts that stop at 9	14	69	38
Counts that stop at 0	10	4	50
Children who finish at least one count with 9	40	96	83
Children who finish at least one count with 0	20	14	100
Omissions			
Counts including any omission	70	75	67
Counts with omission of entire decade	0	32	41
Children omitting at least one number	100	96	83
Children omitting at least one entire decade	0	46	67
Repetitions			
Counts including any repetitions	92	52	54
Counts with repetitions of entire decades	0	11	0
Children repeating one or more numbers	100	85	100
Children repeating one or more decades	0	15	0
Nonstandard numbers			
Counts including nonstandard numbers	0	29	12
Children who used nonstandard numbers at least once	0	54	33
Nonstandard numbers concatenating decade name with 10, 11, or 12 (e.g., 20-10)	—	83	0
Nonstandard numbers concatenating hundreds name to another number (e.g., 100-200)	—	0	100

^aAll measures are percentages, and should be read as "Percentage of . . ." For example, the first measure should be read as "Percentage of counts that stop at 9." These percentages reflect 4 counts per child.

who did not count as high as 20 did not display any obvious regularities in the points at which they stopped. The four most common stopping points of children within this group were 4, 7, 8, and 13. In contrast, an absolute majority of the counts of children who stopped between 20 and 99 ended in a '9' number. The four most common stopping points were 29, 39, 49, and 59. The stopping points of children whose counts exceeded 99 showed yet a third pattern. Here, many counts ended in '9' but even more in '0.' The four most common stopping points were 100, 120, 109, and 129.

Next we shall consider three departures from the standard counting list: omissions, repetitions, and nonstandard numbers. Our focus will be on three features

nonstandard numbers differed considerably among the three groups: none of the children who did not count as high as 20 used any such number, but a majority of those who counted between 20 and 99 and some of those who counted beyond 100 did. The frequency of nonstandard numbers relative to the total of numbers counted was 1% of the total counts in the 20-99 group and 0% for both the 1-19 and the 100+ groups.

The nonstandard numbers that were used invariably involved concatenations of standard numbers. Those of children who stopped counting between 20 and 99 generally came immediately after a number ending in 9 and themselves began with the same decade name followed by a 10 (e.g., 29, 20-10). These nonstandard strings ended soon after they began; all but one ended by the time the child reached the decade name followed by a "12." Children who counted beyond 100 never produced this type of nonstandard number, but fairly often produced a variant in which they reached 100 and then concatenated hundreds names (e.g., 100-100, 100-200).

e. Stability. Recall that the children counted on four occasions, each separated from the previous one by roughly 10 days. This procedure allowed us to examine the stability of their performance over a 1-month period.

In general, children's counts remained in the same category over successive occasions. Between Sessions 1 and 2, and also between Sessions 2 and 3, 83% of the children remained in the same category; between Sessions 3 and 4, 90% did. Over the month-long period between the first and last counting sessions, 79% of the children remained in the same counting category.

Differences in classifications between sessions were generally in the direction of improved counting. Of the changes observed in adjacent sessions, 74% were toward a higher stopping point; of the changes observed between the first and last session, 89% were toward a higher stopping point. Thus, many of those cases in which counting fell within different categories on different occasions probably reflected improved skill rather than measurement error or random variability in performance.

B. COUNTING ON FROM A POINT BEYOND ONE

The counting of children who did not reach the number 20 differed in a large number of respects from the counting of children who proceeded beyond that point. Two basic types of explanations for these differences seemed plausible. One was that the children who did not count as far as 20 were unaware of the generative rule for connecting decade names with digits which applies in the 20s and succeeding decades. The other possibility was that the children knew this rule but had no occasion to use it, since their counts ended before the point at which the rule becomes applicable.

Asking children to count on from various points within and beyond their previously observed counting range was one way to probe the possibility of such unrevealed knowledge. Trials on which children were asked to count on from points *beyond* their previously exhibited counting range seemed likely to be especially revealing. If a child who stopped counting at 12 knew the generative rule for counting in the 20s, then starting the child by saying "21, 22, 23" and then asking him or her to continue seemed likely to reveal the additional knowledge. Similarly, the beyond-counting-range trials seemed likely to reveal whether the counting of children who continued beyond 20 was based on rote learning or on knowledge of an abstract rule. Only mastery of the abstract rule would seem likely to allow children to consistently reach the ends of decades at points beyond those at which they had previously stopped.

1. Method

The same children who participated in the earlier counting experiment, with the exception of three children who had moved or declined to participate, were brought back to the experimental room approximately 2 months after their last spontaneous counting session. First, in order to examine whether the children's counting ability had changed since the earlier assessment, they were asked to count from one, using the same procedure as had been used earlier. Children were divided into the same three expertise groups as previously—0-19, 20-99, and 100+—on the basis of their previous classifications or on the basis of their new count, whichever led to the higher classification. On the basis of their new counts, 6 of the 39 children moved to a higher group, 4 to the 20-99 group and 2 to the 100+ group.

The starting points for the counting-on procedure were varied for the three expertise groups, so as to give children the opportunity to count on both from within and from beyond their previously demonstrated counting range. Those children who had not counted beyond 20 were started successively at 31, 11, 41, and 21. Those who had stopped between 20 and 99 were started successively at 71, 31, 91, and 51. Finally, those who had counted beyond 100 were started successively at 31, 171, 91, and 151. Although the starting points were the same for all children within an expertise group, whether each starting point was within or beyond a given child's counting range was defined individually for that child in terms of his or her previous counts. The criterion for saying that a prompt was beyond the child's counting range was that he or she had not counted as high as the first number of the prompt on any of the five previous counting trials.

Children of all three levels of proficiency were given the following instructions:

Today we're going to play another number game. I'm going to say three numbers. What you need to do is to continue counting on from where I stop. If I say "21, 22, 23," you would say

"24, 25" and so on. I want you to count on as high as you can go. The first number is (e.g.,) 71: 71, 72, 73.

When the child stopped counting, he or she was prompted in the same way as in the counting experiment described above. Once he or she failed to continue, the child was given a running start for the next beginning point.

2. Results

a. 0-19 group. Children who had not previously counted beyond 19 almost never were able to continue much beyond the experimenter's initial prompt (Table II). Not one of the children counted to the next "9." Following the starts in the 20s, 30s, and 40s, none of the children counted beyond the following "5." Thus, these children did not display knowledge of counting beyond what they had exhibited in their previous counts.

b. 20-99 group. In contrast, children who had stopped between 20 and 99 almost always reached the following "9." They did so as often when the following "9" number was beyond as when it was within their previously demonstrated counting range. However, they rarely proceeded beyond that "9" number. Thus, children in the 20-99 group demonstrated substantial ability to complete the decade in which they were started, regardless of whether it was within their

TABLE II
Data on Counting On

Criterion ^a	Group	
	1-19 (N = 6)	20-99 (N = 25)
Criterion digit within child's demonstrated counting competence ^b		100+ (N = 8)
Counts on which child reached next "9" after starting point	0	75
Counts on which child reached next "0" after starting point	0	17
Criterion digit beyond child's demonstrated counting competence		
Counts on which child reached next "9" after starting point	0	90
Counts on which child reached next "0" after starting point	0	6
		31

^aAll data are given as percentages. Criteria should be read as "Percentage of . . ."

^bWithin child's counting competence means that the child has previously counted from one to at least the point of the criterion digit.

previously demonstrated counting competence, but did not exhibit additional knowledge of decade connections.

c. 100+ group. Children who had counted beyond 100 showed yet a third pattern. They always completed the initial decade, usually went on to the next decade when it was within their counting range, and at times went on to the next decade even when it was beyond any of their previous counts. In short, they demonstrated considerable knowledge of the within-decade structure and some knowledge of between-decade connections even when these were beyond their previously demonstrated counting range.

C. MODELS OF THREE LEVELS OF COUNTING EXPERTISE

These data reveal a number of characteristics of children's counting for which any developmental model would need to account. Children who did not count as high as 20 showed no obvious pattern in their stopping points, often omitted or repeated one or a few numbers, did not omit or repeat whole decades, used no nonstandard numbers, and did not count on to the end of the decade when given a three-number running start. Children whose counts ended between 20 and 99 usually stopped at a number ending in "9," often omitted or repeated entire decades, used relatively many nonstandard numbers, most of which involved concatenating decade names with 10, 11, or 12, and counted on to the ends of decades both within and beyond their counting range. Finally, children who counted beyond 100 generally ended their counts at a "9" or at a "0," occasionally skipped whole decades, occasionally formed nonstandard numbers by concatenating hundreds names, and, when asked to count on, both completed decades and went on to later decades within and beyond their previously exhibited counting range.

Before we describe models to account for each of these three counting patterns, it may be desirable to consider some of the factors that would inherently influence the way in which English-speaking children learn to count. Of particular importance are the information that is present in the counting string and the process through which children most likely acquire that information. The sequential nature of counting seems to us critical in determining the acquisition process. Since counting is a sequential activity, children may not grasp the structure inherent in later parts of the number string until their mastery of earlier parts is fairly complete. Thus, children who do not count as high as 20 may not know that there is any structure to the number string because there is little obvious structure in the part of the string that they do know. Children who count even a few numbers beyond 20 could notice that there is a definite structure beyond that point, involving two special sets of numbers and a generative rule connecting them. The generative rule states that whenever a number from one set (i.e., decade names from 20 to 90) is used or heard, successive numbers can be

generated by combining that number with members of a second set (i.e., digit names between 1 and 9) until the numbers in the second set are exhausted. Children who count beyond 100 could note an additional rule involving a third set of numbers (i.e., hundreds names between 100 and 900). In this rule, whenever a member of the third set is used, successive numbers can be created by saying the member of the third set before using the procedure for generating numbers below 100.

This general description suggests several types of knowledge that children would need to count beyond 100: knowledge of the memberships of the three special sets, knowledge of the two generative rules, and knowledge of the "next" relations specifying connections that are not predicted by the generative rules. At this level of description, counting would seem to be a fairly simple skill. In attempting to build models that generate counting performances of the types we observed, however, the complexity of the knowledge underlying counting becomes apparent.

Three models, one corresponding to each level of counting expertise, are shown in Fig. 3. Each of these illustrates representations and processes capable of generating a distinct pattern of counting and counting on. Note that certain parameters of the models will vary for individuals within each group. One child might know that 15 follows 14, while another might think that 16 follows 14; one child might know that 40 follows 39, while another might not; and so on. Thus, the models do not yield completely determinate predictions of individual performance. However, they do yield specific predictions at the level of groups, and the particulars of each model could be filled in to produce separate, determinate predictions for each child. As will be seen, the models are sufficient to account for the main features that differentiate among the three groups of children.

Model I depicts the knowledge hypothesized to underlie counts that ended before 20 (Fig. 3A). This model resembles that of Greeno and co-workers in that the representation includes no particular structure beyond "next" connections. The process is also uncomplicated. First a starting point is chosen; unless some specific point is requested, as in the counting-on instructions, children start with "1." They then say the next number if they can recall it, and continue for as long as they have "next" connections. When they reach a point where they do not have a "next" connection, they either arbitrarily choose a number or stop.

As Fig. 3B illustrates, the representation and process used by children who ended between 20 and 99 (Model II) are hypothesized to be considerably more complicated. Within the representation, numbers can be tagged as members of two lists: the digit repetition list and the rule applicability list. The numbers 1 through 9 are the ones that most often will be on the digit repetition list, though some children's lists may be too short or too long. The functions of this list are to indicate which numbers can be connected with decade names and to avoid the need for separate "next" connections between each successive pair of numbers. Decade names starting with 20 can be on the rule applicability list. This list is

less likely than the digit list to be complete for Model II children. The function of the rule applicability list is to indicate the places where the generative rule, involving the concatenation of the decade name with each of the members of the digit list, can be used.⁵

The Model II process, shown on the bottom of Fig. 3B, operates on the representation as follows. (It is crucial in this example to switch back and forth between the text and the diagram in order to understand either.) Assume that a child is asked to count as high as possible and that this request meets the child's conditions for using the counting process. Since the child was not asked to count from any specific number, she or he assigns *N*, the starting point, a value of one. The child says "one." The word "one" does not include any rule applicability list member, so the program branches to the reader's right. The child then chooses two as the next number, since it is at the end of the "next" link from one. The child says "two." Two also does not include a rule applicability list member, so the child retrieves three and says it. The process continues until the child retrieves and says 20. Twenty is a member of the rule applicability list, so that when the question about the rule applicability list member in the number name is posed, the answer is "yes," and the program branches to the reader's left.

The first question in this part of the program is whether there is a digit list member in the number name. The name 20 does not contain any digit list member, so the next number on the digit list is set equal to one. One is retrieved as the next number on the digit list and is concatenated with 20 to produce 21. The child says "21" and returns to the question of whether there is a digit list member in the name. Now there is a digit list member, one. The question of whether the digit equals 9 is raised; since it does not, the child retrieves the next number on the digit list, two, concatenates it with 20 to produce 22, and says "22." This process continues until the child has generated and said 29. Then the digit does equal 9. Since the number 29 has a specific "next" connection, the child says "30" and repeats the generative cycle.⁶ When the child reaches 39,

⁵The decision to call this list the rule applicability list rather than the decade name list may cause some immediate confusion but should lead to greater clarity in the long run. The set of numbers could not accurately be called the decade name list because 10 is a decade name yet cannot be concatenated with digits to form other numbers. What distinguishes 20, 30, and successive decade names is precisely that with them, the generative rule is *applicable*—hence the name "rule applicability list members."

⁶The decision to connect each member of the rule applicability list to the preceding "9" number rather than to the preceding rule applicability list member was not arbitrary. We tested several 5-year-olds' ability to count by 10s, a skill that they presumably would have if the decade names were linked together. Specifically, the experimenter asked children to count by 10s, prompted them by saying "10, 20," and then asked them to continue. None of the children was able to count in this way, although all had previously counted by ones beyond 30. Thus, it seemed likely that the children's interdecade connections were between "9" numbers and the next rule applicability list member, rather than between successive rule applicability list members.

A

REPRESENTATION:

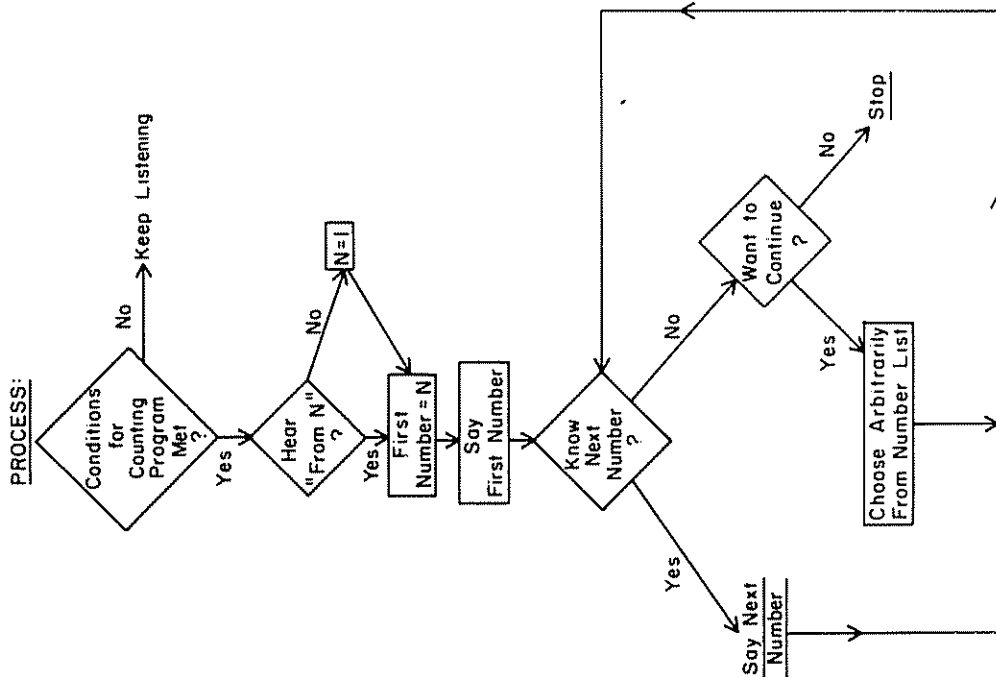
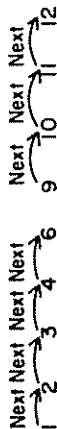
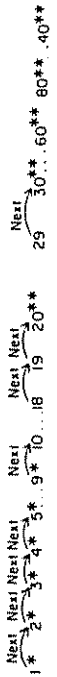


Fig. 3(A-C). Models of counting and counting on. (A) Model 1. (Note: The particular connections and list memberships included here and throughout Fig. 3 representations are chosen purely for purposes of illustration. Also, in the process, N is defined as any integer.)

B

REPRESENTATION:



*ON DIGIT LIST *ON RULE APPLICABILITY LIST

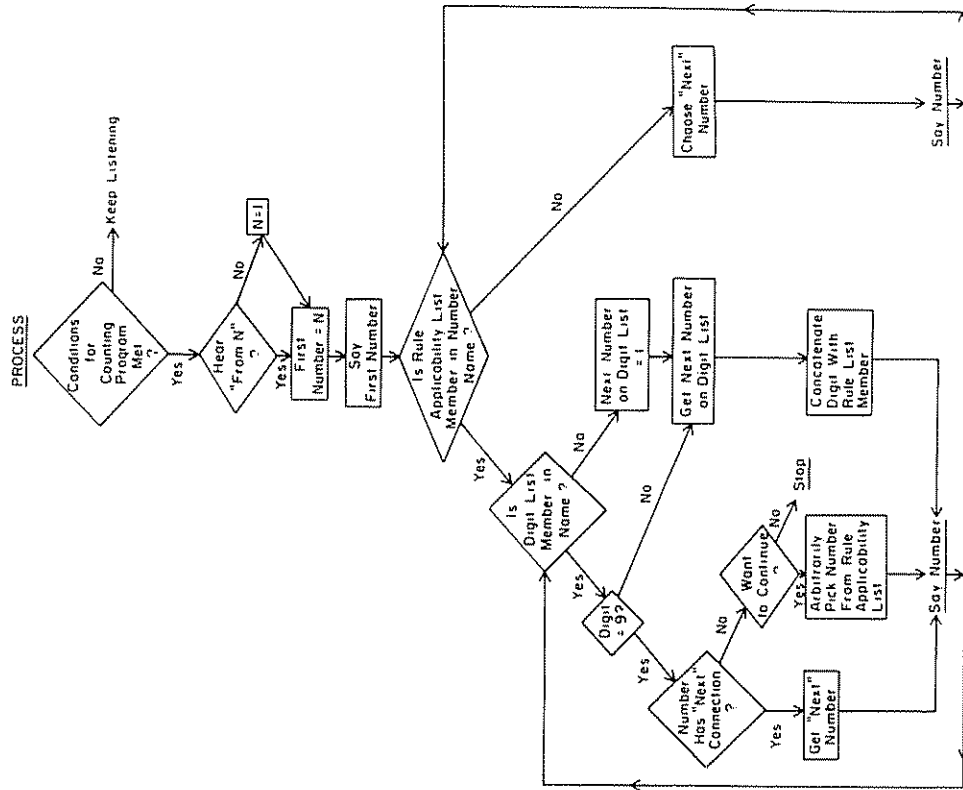


Fig. 3(B). Model II. See next page for Fig. 3(C).

however, there is no specific "next" connection. Therefore, the child reaches an important choice point that accounts for many of the distinctive phenomena in this group's counting data. The child must decide whether or not to continue. A decision not to continue will result in a stop at a "9" number, in this case 39. A decision to continue will result in an arbitrary choice of a number from the child's rule applicability list. This could result in an omission if the number chosen is too far advanced (e.g., 39, 80), in a repetition if the number chosen is not far enough advanced (e.g., 39, 20), or, fortuitously, in the choice of the correct answer (e.g., 39, 40). The process continues until the child reaches a "9" number without a "next" connection and decides to stop.

The Model II representation incorporates only two changes from that of Model I: the addition of the hundreds list, and the completion of the rule applicability and digit lists. The Model III process also works in much the same way as that of the previous model. It proceeds identically until the number 100 is reached. Then, the children note that there is a hundreds list member in the number. They ignore the hundreds list member for purposes of forming subsequent numbers, but remember to add it for purposes of saying them. Remembering to do this may be of more than trivial difficulty; young children in our sample who could count beyond 100 often momentarily forgot the hundreds term at one or more points in their counts (e.g., "124, 125, 26, uh, I mean 126"). This is one reason why within Model III, the instruction to remember to say the hundreds name is separated from the point at which children actually say it.

Model III children can continue counting for a very long time. Such factors as fatigue and boredom seemed as likely as lack of knowledge to determine their stopping points. Therefore, Model III includes a slightly altered stopping rule. When children's desire to continue becomes sufficiently slight, they stop at the next point at which the job would be complete—either at the next "9" or at the next "0" (e.g., at 129 or 130).

These three models account for the major features of the counting of children who stopped within the corresponding ranges. First, let us consider stopping points. Model I does not imply any particular distribution of stopping points, since there is no obvious way of predicting which particular "next" connections will be missing. Model II predicts that stops will occur at numbers ending in "9." Children stop because the number ending in "9" does not have a "next" connection. Model III suggests that stops should occur at either a "0" or a "9," points at which the job is complete.

Now let us consider omissions. Model I children's omissions could come from two sources. First, within the Model I representation, children might have an incorrect "next" connection; illustratively, Fig. 3A indicates that they would say "6" after "4." Second, within the Model I process, children can arbitrarily choose successor numbers when no specific "next" connection exists; within Fig. 3A, this could lead to their skipping from 6 to any other number that they

C

REPRESENTATION:

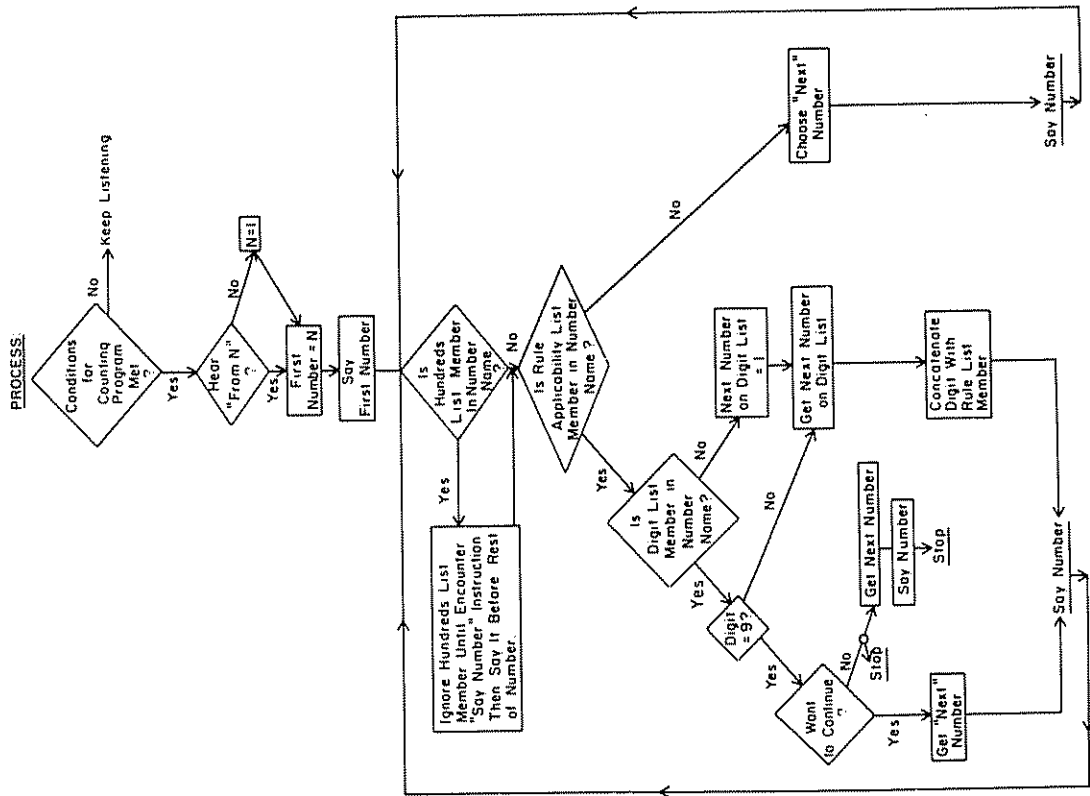


Fig. 3(C): Model III.

knew. Model I does not directly predict the lengths of the omissions that would arise from these mechanisms. However, as exemplified by the particular incorrect link within the Fig. 3A representation, omissions arising from the first source might be expected to be brief, due to children tending to form "next" connections between numbers that were close together in the strings of other people whom they had heard counting.

Model II predicts that brief omissions should rarely occur (there is still the possibility of an incorrect "next" connection) but instead predicts the omission of entire decades. Such omissions would occur when children reach a number ending in 9 that does not have a "next" connection and arbitrarily pick a higher number from the rule applicability list. Within Fig. 3B, children might reach 39, not have a "next" connection, and arbitrarily skip to 60 or 80. Model III predicts no omissions, short or long.

The predictions of the three models concerning repetitions are similar to those concerning omissions, not surprisingly since they are hypothesized to be produced by the same mechanisms. Model I could produce repetitions through two mechanisms: incorrect "next" connections, and arbitrary choices of numbers when no "next" connection existed. Model II could produce repetitions of entire decades following numbers ending in "9" for which the child did not possess a "next" connection. Model III would produce no repetitions of any length.

The frequency and types of nonstandard numbers produced by the three models would also differ. Children who used Model I would not be expected to generate nonstandard numbers. Model II children who had just begun to use the digit repetition list would produce such numbers if the boundaries of their digit lists were too high.⁷ For example, rather than the digit repetition list ending at 9, as in Fig. 3B, it might end at 11. If this were the case, we might expect children's nonstandard numbers to be relatively brief continuations of the standard digit repetition list (it seems unlikely that their boundaries would be off by very much). Children who used Model III would presumably have overlearned the digit list, but might have difficulty in remembering the recursive procedure for forming numbers with hundreds list members; this could lead to nonstandard concatenations of hundreds list members.

Finally, consider the counting-on data. Model I includes no mechanism for children to count on if they are started beyond their counting range; they would lack the necessary "next" connections. Children using Model II would be expected to be able to count on to the end of the decade from any point that was

⁷Cases in which the child's digit repetition list did not extend quite far enough may well have occurred, but were difficult to detect from the production data. The most direct data concerning this possibility were the relatively large numbers of stops at 26, 27, and 28 shown in Fig. 2. The children who stopped at these points often did not proceed beyond 29 on any of their counts and had a higher than average frequency of nonstandard numbers. These findings were consistent with the hypothesis that these children were still in the process of mastering the digit repetition list.

recognized to include a rule applicability list member and to continue to the next decade if the "next" connection was known. Even if a child using Model II did not yet know that a particular number beyond his or her counting range was on the rule applicability list, he or she might infer its status from hearing it concatenated with one, two, and three successively (in the same way that an adult who heard someone say "blypty-one, blypty-two, blypty-three" would probably continue to blypty-nine). Children who used Model III would be expected to perform similarly to those who used Model II.

Almost all of these predictions of the models fit the children's counting. Only one observation proved recalcitrant: the failure of Model II children in the counting-on experiment to count beyond the end of the decade within which they were started, even when their previous counting from 1 demonstrated knowledge of the relevant "next" connection. Illustratively, 7 children who had previously counted flawlessly to 49 or higher did not progress beyond 39 when asked to count on from 31. This is not an easy finding to explain. The best that we can do is to place it in the general framework of overlearning phenomena. It seems almost certain that the "next" relations connecting digit repetition list members with each other are overlearned to a much larger degree than those connecting "9" numbers with rule applicability list members. Such overlearning may enable young children to access the "next" relations on the digit repetition list even when they have not found their place in the counting sequence; they may approach the task as if the decade name were a nonsense syllable (i.e., the blypty-one, blypty-two example). Admittedly, more research is needed to explain the phenomenon (cf. Fuson & Richards, in press, for a similar finding on a "What is the next number?" task).

With this single exception, the Fig. 3 models were consistent with the major features of the data. The distributions of stopping points, omissions, repetitions, and nonstandard numbers were what the models would predict, as were all but the one finding on children's counting on within and between decades.

IV. Preschoolers' Knowledge of Numerical Magnitudes

A. EXISTING RESEARCH ON ADULTS AND CHILDREN

A second central aspect of children's knowledge of numbers is their knowledge of numerical magnitudes. Research in this area was greatly stimulated by the report by Moyer and Landauer (1967) that comparing the magnitudes of digits of discrepant sizes (e.g., 4 and 8) took less time than comparing the magnitudes of digits of more comparable sizes (e.g., 4 and 5). At the time, this symbolic distance effect must have appeared quite surprising, because the exper-

iment was replicated at least five times within the next 5 years (Aiken & Williams, 1968; Fairbank, 1969; Knoll, cited in Moyer & Bayer, 1976; Parkman, 1971; Sekuler, Armstrong, & Rubin, 1971).

The symbolic distance effect was especially intriguing because several of the most obvious models of numerical comparison were directly contradicted by the data. One way that numerical comparison problems could be solved would be to count from 1 until encountering one of the numbers; that number would be the smaller, and the remaining number would be the larger (Parkman, 1971). Such a model does not account for the distance effect, however; it predicts that solution time should be a function of the size of the minimum number, but not of the split between the two numbers. Another way to solve magnitude comparison problems would be to count on from one number until the other was found; if it were not found in some period of time, the count would be initiated from the other number. This model would predict a reverse symbolic distance effect, however; the closer the two numbers, the sooner the second number should be encountered. A third possibility, direct associative retrieval, would seem to yield equal solution times for all comparisons, again different from the observed pattern.

As Moyer and Landauer (1967) pointed out, however, the psychophysics literature provides a well-documented analog to the symbolic distance effect. Psychophysicists long ago noted that the more discrepant on some dimension the magnitudes of two visible or audible stimuli, the faster they can be compared. Another well-known psychophysical phenomenon also proved to be present in the abstract comparisons; the smaller the size of the smaller stimulus, the quicker is the judgment (e.g., 2 and 4 can be compared more quickly than 4 and 6). This has been labeled the min effect.

To account for these parallels, several investigators have hypothesized that physically and symbolically presented magnitudes share a common underlying representation (Banks, 1977; Moyer & Dumais, 1978; Shepard & Podgorny, 1978). For numerosity, this representation is assumed to resemble a logarithmic scale, with the values at the high end of the scale more densely packed than the values at the low end. The symbolic distance effect is said to arise from the fact that it is easier to discriminate between points that are far apart in the representation; the min effect is attributed to representations of magnitudes of the same linear disparity being farther apart at the low than at the high end of the scale. Multidimensional scaling of the results of numerical comparison tasks (Shepard, Kilpatrick, & Cunningham, 1975), random number production tasks (Banks & Hill, 1974), and subjective magnitude judgment tasks (Rule, 1969) have provided converging evidence for this view of the representation of numerical magnitudes.

Disagreements have arisen over several other characteristics of the representation, however. Two broad classes of models have emerged: analog and discrete. The central assumptions underlying analog models are that representations of all

types of magnitudes preserve continuous information about physical size, and that these analog values are compared directly, perhaps through some type of random-walk process (e.g., Moyer & Dumais, 1978; Pollock, 1980). In contrast, the central assumption underlying discrete models is that magnitudes are grouped into categories, each of which carries a semantic code (e.g., large, small) and that it is the codes which are compared. Illustratively, in one discrete model, Banks, Fujii, and Kayra-Stuart (1976) suggested that each digit is associated with the code "big" with a probability proportional to the logarithm of that number, and with the code "small" with a probability of one minus its logarithm. When asked to compare numerical magnitudes, people generate codes for the two numbers, answer if the codes discriminate between the numbers, and, if they do not, regenerate the codes until they do discriminate between them. Substantial bodies of data have been collected supporting each type of model, though it is our impression that the discrete models have been most consistently in accord with the data on *numerical* magnitude comparisons.

One limitation of both the analog and the discrete models is that their proponents fail to specify whether they are intended as characterizations of long term memory contents or as characterizations of temporary data structures formed for the purpose of performing a particular experimental task. The models are presented as if the analog or discrete representations were retrieved directly from long-term memory, but this view seems implausible. Adults can compare the magnitudes of a vast set of numbers, yet almost certainly do not possess in long-term memory either analog values or categorical labels corresponding to each number that they know. Interestingly, the force of this reservation may not be nearly as great with preschoolers, a population to whom the models have not been applied, as with adults. In a recent experiment (Siegler & Robinson, 1981) we found that preschoolers most often label 5 as a medium-size number and 9 as a big number, regardless of whether the context is the numbers 1-9 or the numbers 1-1 trillion. Adults also label 5 as medium size and 9 as big in the 1-9 context, but, not surprisingly, refer to both as small numbers in the 1-1 trillion context. Thus, although both analog and discrete models probably reflect temporary data structures in adults, they more plausibly reflect long-term memory representations as well as temporary data structures in very young children.

As alluded to in the paragraph above, despite the large amount of work on adults' understanding of numerical magnitudes, very little is known about how such understanding develops. We were able to locate only two published studies involving young children: Sekuler and Mierkiewicz (1977) and Schaeffer, Eggleston, and Scott (1974). Unfortunately, each of these studies has important limitations that prevent us from drawing strong conclusions from them.

Sekuler and Mierkiewicz (1977) presented magnitude comparison problems involving the digits 1-9, with equal numbers of items having each level of split from 1-8. They found that the solution latencies of children as young as 5 years

showed an effect of symbolic distance. However, their analyses averaged items over values of min, and therefore did not distinguish between the effects of minima and those of symbolic distances, a serious defect since the two variables were highly correlated within their problem set. This omission would be unfortunate with adults, where the absolute size of the numbers has been found to exercise an effect independent of symbolic distance, but seems even more serious with young children, whose knowledge of small and large numbers has been found to differ greatly in many situations (e.g., Fuson & Richards, 1982; Gelman & Gallistel, 1978; Ginsburg, 1977; Siegler, 1981a).

Schaeffer, Eggleston, and Scott's experiment exhibits different but equally noteworthy problems. They presented 2- to 5-year-olds' numerical magnitude comparisons involving the digits 1-9 and having a split of either 1 or 4 digits. They interpreted their results as indicating that children divide numbers into small and large categories and that they learn to perform numerical comparisons in the order: (1) large distance, with numbers falling into different categories; (2) large distance, with both numbers in the same category; (3) small distance, with both numbers within the small number category; and (4) small distance, with both numbers within the large number category.

It is difficult to know how to evaluate this account. Schaeffer *et al.* (1974) did not describe the data that led them to believe that young children divided numbers into small and large categories. Their classification criteria were inconsistent; in comparisons involving large distances, the boundary between the small and large categories was placed at 5, but in comparisons involving small distances, it was placed at 4. Further, it is not clear why crossing a category boundary was essential in determining the difficulty of comparisons involving large distances (Stage 1 versus Stage 2) but not in determining the difficulty of comparisons involving small distances (Stage 3 versus Stage 4).

Thus, at present, we know little about how understandings of numerical magnitudes develop. This lack of knowledge is unfortunate, not only because of the inherent place of such knowledge within children's understandings about numbers, but also because developmental research would seem to have substantial potential to illuminate some of the general issues that have arisen in the magnitude comparison literature. As Banks (1977) pointed out, almost all magnitude comparison studies have involved one of two types of stimulus materials: overlearned material, such as letters and digits, and arbitrary material, such as different-colored sticks or nonsense syllables. Neither of these types of material seems ideally suited to studying acquisition processes. In the first case, acquisition is already complete; in the second, the material being acquired is inherently unrepresentative of the semantically rich material that people learn about in the world outside the laboratory. In contrast, studying children's acquisition of the ordering of real-world materials such as numbers would seem to overcome both of the problems. The materials are semantically rich (Lehman, 1979), yet can be studied at a point where they have not been entirely mastered.

The purpose of our first numerical comparison experiment was therefore to obtain a more detailed picture of knowledge of numerical magnitudes at ages where such knowledge is sufficient to produce above-chance comparison performance but insufficient to produce consistently correct performance.

B. PRESCHOOLERS' NUMERICAL MAGNITUDE COMPARISONS

1. Method

The same 39 children who participated in the counting-on experiment were brought back to the experimental room and given the following instructions:

Today we're going to help Bunny and Monkey (2 stuffed animals that were present) to play a card game. I'm going to hold up 2 cards so that I can see the numbers. I'll tell you what the numbers are, and you need to tell me whether Bunny's number is more or whether Monkey's number is more. OK, Bunny has 7 and Monkey has 5. Which one is more: 7 or 5?

Similar questions were asked for all 36 nonidentical pairs of the digits 1-9. If a child's interest appeared to be flagging, the session was terminated for that day and continued the next. Children were periodically told that they were doing well, regardless of their actual level of performance.

Each participant was presented the group of 36 items on four occasions. The items were presented in one randomly generated order on the first and fourth occasions and in the reverse order on the second and third occasions. The order of mention of the numbers within each problem (e.g., "Which is bigger, 9 or 5?" versus "Which is bigger, 5 or 9?") was reversed on the second and fourth trials from the order on the first and third trials. The four presentations occurred at roughly 10-day intervals. Thus, as in the first counting experiment, the testing period lasted about 1 month.

2. Results

a. Absolute level of performance. The period from 3 to 5 years proved to be one in which considerable development occurred in children's ability to compare digit magnitudes. Three-year-olds were correct on 56% of their comparisons, 4-year-olds on 81%, and 5-year-olds on 90%. In order to obtain an unbiased estimate of the children's knowledge, the effect of chance was corrected by the formula

$$\text{knowledge} = \frac{(\text{observed percentage correct} - \text{chance percentage correct})}{(1 - \text{chance percentage correct})}$$

Since the chance percentage correct was 50%, inserting into the formula the above data on observed percentage correct indicated that 3-year-olds knew 12% of the comparisons, 4-year-olds 62%, and 5-year-olds 80%.

TABLE III
Percentage of Variance Accounted for by Alternative Predictors of Magnitude Comparison Errors

Age of child	Model			Banks [A (log(min)) + B (1 - log(max))]
	Linear (min/split)	Welford function [log (max/split)]	Parkman (min)	
5-year-olds	52	55	48	41
4-year-olds	69	70	42	55

b. Predictors of performance. We next applied the same types of regression analyses to the young children's error patterns as previously have been employed in examining adults' reaction times. The variables that were used to predict the percentages of errors on the 36 problems were the size of the minimum number, the size of the maximum number, the distance between the two numbers, and the sum of the numbers. None of these variables proved significantly predictive of 3-year-olds' performance. In contrast, both the size of the minimum number and the distance between the numbers proved predictive of both 4- and 5-year-olds' behavior. In the case of the 4-year-olds, the two variables accounted for 69% of the variance in the number of errors on the 36 problems, while in the case of the 5-year-olds, they accounted for 52%. For each age group, the size of the minimum number accounted for the larger percentage of variance: for the 4-year-olds, $r_{\text{min-split}} = -.68$, while $r_{\text{split-min}} = .48$; for the 5-year-olds, $r_{\text{min-split}} = -.59$, while $r_{\text{split-min}} = .28$. The linear combination of min and split accounted for at least as much of the variance in children's errors as several nonlinear combinations of the variables that have been proposed as functional models in previous studies; the respective figures are shown in Table III.

c. Consistency of performance. The children's percentages of correct answers showed a fair degree of consistency over the four occasions. For example, on the first and second occasions, the 3-year-olds' performance correlated $r = .43$, the 4-year-olds' performance correlated $r = .79$, and the 5-year-olds' performance correlated $r = .58$. Across all three age groups, the correlations between performance on the two occasions was $r = .79$.

The data on consistency of performance on the same problems over different occasions were also revealing. Three-year-olds were correct on all four occasions on 10% of problems, correct on some but not all occasions on 87%, and incorrect on all occasions on 3%. The corresponding percentages for 4-year-olds were 54, 45, and 1%, respectively, and for 5-year-olds 67, 33, and 0%, respectively. The lack of systematically incorrect responding was striking for all three age groups. It suggested that any model of young children's numerical magnitude comparisons could not be completely determinate in its predictions of the prob-

lems on which each child would err; rather, the model would need to include one or more probabilistic processes that would sometimes but not always lead to errors on given problems.

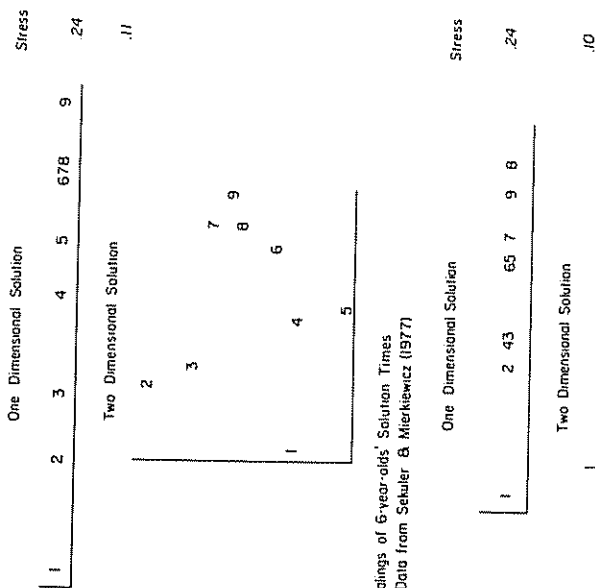
d. Representations of numerical magnitudes. What representations of numerical magnitudes might account for the obtained error patterns? One method that can be used to address such questions is multidimensional scaling. As input, multidimensional scaling routines take data concerning similarities or proximities among the possible pairs of stimuli. The percentage of errors can be used as one index of similarity, the assumption being that errors reflect the difficulty of discriminating the points in some type of representational space. As output, multidimensional scaling routines produce spatial arrangements of the individual stimuli that minimize stress, stress being a badness of fit measure.

The particular multidimensional scaling procedure that we used to examine children's magnitude representations was the nonmetric version of KYST, with stress formula 1 and the primary method for dealing with ties. In line with Kruskal and Wish's (1978) suggestion, we limited our consideration of possible dimensionalities to 2, since 9 stimuli were being scaled. The input to the scaling algorithm was the percentage of errors that children made in comparing each of the 36 possible pairs of digits from 1 to 9; the output was an arrangement of the 9 digits in a one- or two-dimensional representational space.

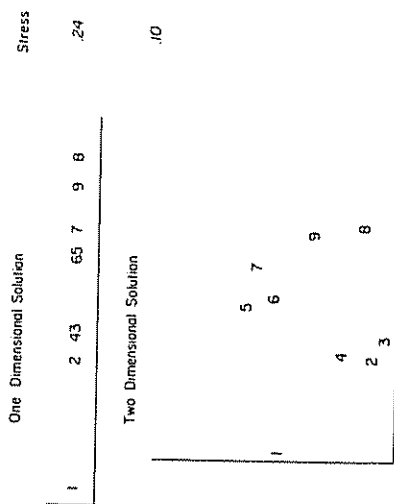
The most striking result of the multidimensional scalings was that the numbers did not fit especially well into the compressive logarithmic function generally believed to characterize representations of numerical magnitudes. Rather, they seemed to fall into clusters, with quite small distances within clusters and quite large distances between them. Consider, for example, the data of the 4-year-olds shown in Fig. 4A. There are some reasons for preferring the two-dimensional representation over the one-dimensional one—notably that the stress declines from .24 to .11—but even the one-dimensional representation does not particularly closely resemble the hypothesized logarithmic representation. Rather, the numbers seem to arrange themselves into four clusters: (1), (2, 3), (4, 5), and (6, 7, 8, 9).

This finding motivated us to reanalyze the reaction time data of Sekuler and Mierkiewicz (1977) with slightly older children. The scaling results with their 6-year-olds are shown in Fig. 4B. Again, note that the results do not fit especially well the compressive logarithmic function postulated previously; rather, the numbers seem to fall into a few clusters.

In order to objectively test the impression of clustering in our own error data, we next performed hierarchical clustering analyses (using the diameter method; see Johnson, 1967). The results for the 4-year-olds are shown in Fig. 4C. As can be seen in the patterns of ovals (stimuli within the smallest number of ovals are the most similar), the clusters are quite similar to those revealed impressionistically.



B. Scalings of 6-year-olds' Solution Times (Data from Sekuler & Mierkiewicz (1977))



C. Hierarchical Clustering of 4-Year-Olds' Errors (Our Data)

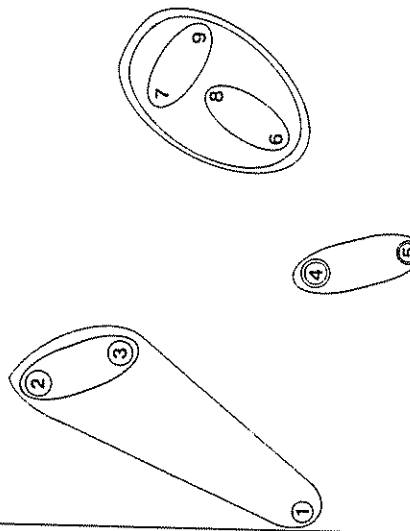


Fig. 4. Multidimensional scaling solutions for magnitude comparison data: (A) scalings of 4-year-olds' errors (our data); (B) scalings of 6-year-olds' solution times (data from Sekuler & Mierkiewicz, 1977); (C) hierarchical clustering of 4-year-olds' errors (our data).

The emergence of two distinct dimensions in the multidimensional scaling solutions implied that relative difficulty of the magnitude comparison problems was determined by the two numbers being in the same or in different clusters, rather than by any differential distance between clusters. In order to objectively test this hypothesis within our own data, a regression analysis was performed using between- versus within-cluster status as a predictor of the relative difficulty of the 36 comparison items. The cluster membership variable alone accounted for 63% of the total variance in the number of errors, not that different from the 73% that could be accounted for using the absolute distances between digits in the scaling solution as the predictor. We concluded from these data that within-versus between-cluster status of the comparison items was a major determinant of problem difficulty.

Examination of the scaling and clustering analyses also suggested possible explanations for the min and split effects. In all of the analyses, the digit "1" was in a cluster by itself; this placement would contribute to both min and split effects, since comparisons involving 1 have the lowest possible minimum and on average will have the largest possible splits. Also, in all of the scalings the number of numbers within the clusters at the small end of the scale is smaller than the number of numbers in the clusters at the large end; again, this distribution suggests a greater probability of the relatively easy between-category comparisons when mins are small and splits are large.

Thus, several sources of evidence were consistent with the notion that young children's representations of numerical magnitudes were neither undifferentiated nor closely correspondent to a logarithmic function. Rather, the digits seemed to be represented in terms of a small number of clusters. It must be noted, though, that all of these sources of evidence involved the end products of magnitude comparisons. The children's categorizations of the numbers could be inferred only from their overall performance on the magnitude comparison task, a performance that was produced by a process as well as a representation. In order to obtain more direct evidence about children's representations of numerical magnitudes, we therefore thought it important to determine (1) whether children's categorizations of the digits, measured independently of their magnitude comparisons, were related to their magnitude comparison performance; and (2) if the relationship was present, whether teaching children a new clustering scheme would influence their pattern of correct answers and errors in comparing magnitudes. The two experiments reported below were addressed to these questions.

C. VERBAL LABELING OF NUMBERS

In both Trabasso's (1977) and Banks, Fujii, and Kayra-Stuart's (1976) discrete category models, people were said to attach a semantic code to each stimulus being compared. The only empirical support for this claim, however, was that it was consistent with the semantic congruity effect. In the experiment presented

here, we attempted to study more directly children's semantic coding by examining the verbal labels that they assigned to numbers. In particular, we were interested in (1) whether the labels children assigned to numbers could potentially have been useful if they applied them in comparing numerical magnitudes; and (2) whether the quality of children's labeling, considered across the digits, was predictive of performance on the magnitude comparison task.

1. Method

Children were brought back to the experimental room roughly a week after their last magnitude comparison session and given the following instructions:

Today I'm going to ask you some questions about the numbers from 1 to 9. Some of these numbers are big numbers, some are little numbers, and some are medium numbers. I'm going to say a number, and you need to tell me if the number is a big number, a little number, or a medium number.

Then the experimenter said a number and asked, "OK, is N a big number, a little number, or a medium size number?" The numbers were arranged in random order, save for the restriction that each third of the presentation order needed to include one number from among the digits 1-3, one from among the digits 4-6, and one from among the digits 7-9.

2. Results

a. Absolute level of performance. One concern that we had before conducting this experiment was that the children would not understand the task. This fear proved groundless. Even the 3-year-olds demonstrated that they understood what they were being asked to do. The mean value of the numbers that they termed small was 3.14, of the numbers that they labeled medium 5.15, and of the numbers that they labeled big 6.00. More 3-year-olds labeled each of the numbers 1, 2, and 3 as small than labeled them as big; more of them labeled 7, 8, and 9 as big than labeled them small. Thus, even the youngest children's labeling demonstrated some knowledge about the magnitudes of the numbers.

The 3-year-olds' relatively accurate labeling does not imply that there was no development in skill at applying the labels. Older children's labelings were much more differentiated in terms of the numbers they included. By age 5, the mean values of numbers assigned the small, medium, and large labels were 2.16, 5.12, and 7.75, respectively. Illustratively, 9 of the 10 5-year-olds labeled the number 1 "small," 9 of the 10 labeled the number 5 "medium," and all 10 labeled the number 9 "big."

b. Potential usefulness of labels. The first main issue addressed in this experiment was whether children's labeling of the numbers could have been

useful on the magnitude comparison task. In order to examine this issue, we needed to quantify the implications of the labels for the magnitude comparisons. Suppose that a child used *only* his or her labels in comparing numerical magnitudes. Such a procedure could lead to correct answers (e.g., if 7 was labeled "big" and 5 was labeled "little"), to incorrect answers (e.g., if 7 was labeled "medium" and 5 was labeled "big"), or to indeterminately correct answers (e.g., if both 7 and 5 were labeled "medium"). Therefore, as a first rough estimate, the quality of each child's labels was quantified by examining the labels assigned to each of the 36 possible pairs of digits and assigning a score of 1 on pairs on which the labels would lead to the correct answer, a score of 0 to comparisons on which the labels would lead to an incorrect answer, and a score of .5 on comparisons on which the labels would lead to an indeterminately correct outcome.⁸ Under this scoring system, a child who assigned labels to numbers haphazardly would be expected to answer correctly on 50% of the problems, as would a child who assigned the same label to all 9 digits. The best possible score would be produced by a child who labeled 1, 2, and 3 as small numbers, 4, 5, and 6 as medium numbers, and 7, 8, and 9 as big numbers; this strategy would produce an expected value of 88% correct performance.

This measure indicated considerable improvement with age in the quality of children's labeling. The 3-year-olds' labeling would have led to an average of 58% correct answers, the 4-year-olds' to 75%, and the 5-year-olds' to 80%. Thus, the 4- and 5-year-olds' labels were potentially useful on the numerical comparison task; if the children had used them, they would have performed at a level well above chance.

The number of correct answers predicted by the above measure of the quality of each child's labels also proved to be closely related to the number of correct answers the child had produced on the numerical comparison task. Across the three age groups, the correlation was $r = .80$. The correlations were also substantial within each of the three age groups; $r = .73$ for the 3-year-olds, $r = .55$ for the 4-year-olds, and $r = .64$ for the 5-year-olds.

In addition to the labels predicting children's number of correct answers on the magnitude comparison task, they also predicted the relative difficulty of the individual problems. Consider the results with the 4-year-olds. First, we calculated the probabilities, over all of the 4-year-olds, that each label was assigned to each number. The probabilities of each possible pair of labels being assigned to

⁸The estimate of .5 as a child's probability of correctly answering a problem on which he or she assigned identical labels to the two numbers was chosen as a first approximation, in lieu of a specific model of the comparison process. The magnitude comparison model that was eventually hypothesized (described below) actually reaches a somewhat different estimate of the probability of a correct answer in such cases, but for reasons that could not be explained at this point. None of the data reported in this section is importantly affected by the differences between the two procedures for estimating the probabilities.

each pair of digits were then calculated to yield the values within the equation:

$$p \text{ (correct on each magnitude comparison problem)} = \frac{p \text{ (correctly ordered labels)}}{p \text{ (correctly ordered labels)} + p \text{ (incorrectly ordered labels)}}$$

An illustration of the workings of this equation may be helpful. If the number 5 was labeled little, medium, and big with probabilities of .2, .5, and .3, respectively, and if the number 7 was labeled little, medium, and big with probabilities of .1, .5, and .4, respectively, then the probability of correctly ordered labels would be $.38 [(.5 \times .2) + (.4 \times .5)]$, the probability of incorrectly ordered labels would be $.23 [(.5 \times .1) + (.3 \times .5)]$, and the probability of a correct magnitude comparison would therefore be $.62 (.38/.61)$.

The results of this equation were used to predict the number of errors that the 4-year-olds previously had made on the 28 magnitude comparison problems that did not involve the number 1 (problems involving 1 were excluded because the Fig. 4 clustering results led us to believe that 1 had a separate "smallest number" label that was not tapped by our possible labels of little, medium, and big). Using the above scoring procedure, the 4-year-olds' probability of labeling each digit as small, medium, or large was found to account for 66% of the variance in their number of errors on the magnitude comparison problems.

Thus, to summarize, the labeling experiment demonstrated that the labels children applied to numbers could logically have led to above-chance performance in the magnitude comparison context, that the labels were in fact correlated with individual differences in the percentage of correct answers on the comparison task, and that the labels predicted the relative difficulty of the magnitude comparison problems.

A central question remained, however: "Was the children's labeling of the numbers used in the process by which they compared magnitudes, or did performance on the one task simply predict performance on the other?" This issue was addressed in the next experiment.

D. EFFECTS OF TEACHING A LABELING STRATEGY

If children's labels were functionally involved in the magnitude comparison process, then teaching them a specific set of labels for the numbers might be expected to influence their later numerical comparisons. In particular, children who used the labeling scheme that they were taught would be expected to be more successful on between-category comparisons (defined in terms of that labeling scheme) than on within-category ones. A control group that was not taught this particular labeling scheme would not be expected to show as large a between-category versus within-category difference, because they would not be as likely to use this particular set of labels. In addition, if children had previously

selected labels independently for each number, without regard for the other number's label, then instructing them in the use of an entire set of labels would be expected to reduce their percentage of incorrect answers (as will be explained in detail in the models below). The primary purpose of the training experiment was to test these predictions.

1. Method

Twenty children of mean age 4 years, 10 months were recruited from a day care center serving a predominantly middle-class area. These children were randomly assigned to one of two groups: a label training group and a control group. Children were brought individually to a small vacant room in the day care center. Those in the training group were told:

Today we're going to learn about the numbers from 1 to 9. Some of the numbers are little numbers, some are medium numbers, and some are big numbers. I will tell you which numbers are little, which are medium, and which are big. One, two, and three are little numbers; four, five, and six are medium numbers; and seven, eight, and nine are big numbers. Now I'm going to say a number and you need to tell me if the number is a little number, a medium number, or a big number.

The experimenter then said "N: is N a little number, a medium number, or a big number?" If the child's label matched the one the experimenter had provided, the experimenter said "Good, N is a (big) (medium) (little) number." If it did not match, the experimenter said "No, N is a (big) (medium) (little) number." On each trial block, the digits were presented in a different randomly generated order. After the completion of each trial block, the experimenter repeated the initial information about which numbers were big, little, and medium. Then the next trial block was presented. The procedure continued until the child's labeling of all 9 digits matched the experimenter's on 2 consecutive trial blocks, or until 10 trial blocks had been given, whichever came first.

Children in the control group were exposed to one of two procedures. Those in the labeling control were asked on five successive trial blocks to label each of the digits as little, big, or medium. These children's procedure was identical to that received by children in the training group except that they were not instructed to use any particular labels before each trial block and were not given feedback about their performance. Children in the contact control were given no particular experience with labeling of numbers; they simply interacted with the experimenter for a period of time equivalent to children in the label training group.

The next day, children were brought back to the experimental room and presented the 36 standard magnitude comparison problems. One day later, they were presented the same problems but in reverse order.

2. Results

The data of children in the two control groups were indistinguishable; therefore, they were considered together in all subsequent analyses.