Reprinted for private circulation from CHILD DEVELOPMENT MONOGRAPH
Serial 99, 1965, Vol. 30, No. 1

V. ON THE BEHAVIORAL FOUNDATIONS OF MATHEMATICAL CONCEPTS

PATRICK SUPPES
Stanford University

INTRODUCTION

The title of this paper will perhaps mean different things to different people. Philosophers and mathematicians interested in the foundations of mathematics and the philosophy of language may think I intend to pursue a systematic pragmatics built around such notions as Ajdukiewicz' concept of acceptance. Actually, I am going in a different direction. What I want to do is outline present applications of mathematical learning theory to mathematical concept formation. The aims of this paper are primarily constructive, that is, to contribute to the development of a scientific theory of concept formation. Before I turn to this subject, however, I want to comment on two general aspects of the teaching of mathematical concepts.

The first concerns the much-heard remark that the newer revisions of the mathematics curriculum are particularly significant because of the emphasis they place on *understanding* concepts as opposed to the perfection of *rote* skills. My point is not to disagree with this remark, but to urge its essential banality. To understand is a good thing; to possess mere rote skill is a bad thing. The banality arises from not knowing what we mean by *understanding*. This failure is not due to disagreement over whether the test of understanding should be a behavioral one. I am inclined to think that most people concerned with this matter would admit the central relevance of overt behavior as a measure of understanding. The difficulty is, rather, that no one seems to be very clear about the exact specification of the behavior required to exhibit understanding. Moreover, apart even from any behavioral questions, the very notion of understanding seems fundamentally vague and ill defined.

To illustrate what I mean, let us suppose that we can talk about understanding in some general way. Consider now the concept of triangularity.

This research has been supported by the U. S. Office of Education, Department of Health, Education, and Welfare, the National Science Foundation and the Carnegie Corporation of New York.

Does understanding this concept entail the understanding that the sum of the interior angles is 180°, or that triangles are rigid whereas quadrilaterals are not, or the ability to prove that if the bisectors of two angles of a triangle are equal then the triangle is isosceles? This example suggests one classical philosophical response to our query, that is, to understand a concept means, it is said, to know or believe as true a certain set of propositions that use the concept. Unfortunately, this set is badly defined. It is trivial to remark that along these lines we might work out a comparative notion of understanding that is a partial ordering defined in terms of the inclusion relation among sets of propositions that use the concept. Thus, one person understands the concept of triangularity better than a second if the set of propositions that uses the concept and is known to the first person includes the corresponding set for the second person. (Notice that it will not do to say simply that the first person knows more propositions using the concept, for the second person might know fewer propositions but among them might be some of the more profound propositions that are not known by the first person; this situation corresponds to the widely held and probably correct belief that the deepest mathematicians are not necessarily the best mathematical scholars.)

But this partial ordering does not take us very far. A more behavioral line of thought that, at first glance, may seem more promising is the response of the advocates of programmed learning to the charge that the learning of programmed material facilitates rote skills but not genuine understanding of concepts. They assert that if the critics will simply specify the behavior they regard as providing evidence of understanding, the programmers will guarantee to develop and perfect the appropriate repertory of responses. This approach has the practical virtue of sidestepping any complex discussion of understanding and supposes, with considerable correctness, no doubt, that without giving an intellectually exact analysis of what to understand a concept means, we still can obtain a rough consensus at any given time of what body of propositions we expect students to master about a given concept. This is the appropriate practical engineering approach, but it scarcely touches the scientific problem.

In this paper I do not pretend to offer any serious characterization of what it means to understand a concept. I do think that the most promising direction is to develop a psychological theory of concept transfer and generalization. The still relatively primitive state of the theory of the much simpler phenomena of stimulus transfer and generalization do not make me optimistic about the immediate future. For immediate purposes, however, let me sketch in a very rough way how the application of ideas of transfer and generalization can be used to attack the banality mentioned earlier in the standard dichotomy of understanding vs. rote skill.

We would all agree, I think, that such matters as learning to give the multiplication tables quickly and with accuracy are indeed rote skills. But there is also what I consider to be a mistaken tendency to extend the label "rote skill" to many parts of the traditional mathematics curriculum at all levels. The body of mathematical material tested, for example, by the British Sixth Form examinations is sometimes so labeled by advocates of the newer mathematics curriculum. In terms of the accepted notion of rote skill developed and studied by psychologists, this is a mistake, for the production of a correct response on these examinations cannot be explained by any simple principle of stimulus-response association. Moreover, the problems of transfer involved in solving typical British Sixth Form examination problems, in comparison with the kind of examination set by advocates of the newer mathematics curriculum may, in fact, require more transfer of concepts; at least, more transfer in one obvious way of measuring transfer, that is, in terms of the number of hours of training spent in relation to the ability to solve the problems by students matched for general background and ability. I recognize that these are complicated matters and I do not want to pursue them here. Also, I am fully in sympathy with the general objectives of the newer mathematics curriculum. I am simply protesting against some of the remarks about understanding and rote skills that occur in the pedagogical conversations and writings of mathematicians.

The second general point I want to mention briefly is of a similar sort. I have in mind the many current discussions of the efficacy of the discovery method of teaching. Such discussions seem to provide yet one more remarkable example, in the history of education, of a viewpoint achieving prominence without any serious body of empirical evidence to support or refute its advocates. From the standpoint of learning theory, I do not even know of a relatively systematic definition of the discovery method. I do not doubt that some of its advocates are themselves remarkably capable teachers and able to do unusual and startling things with classes of elementary-school children. The intellectual problem, however, is to separate the pedagogical virtuosities of these advocates' personalities from the systematic problem of analyzing the method itself. Workable hypotheses need to be formulated and tested. I know that a standard objection of some advocates of the discovery method is that any quick laboratory examination of this teaching method vs. a more standard immediate reinforcement method, particularly as applied to young children, is bound not to yield an unbiased test. The results and the implications of the methods, it is said, can only be properly evaluated after a long period. I rather doubt that this is the case but, if it is so, or if it is propounded as a working hypothesis by advocates of the method then, it seems to me, it is their intellectual responsibility to formulate proper tests of a sufficiently sustained sort.

I realize that my remarks on this subject have the character of *obiter dicta.* On the other hand, in a more complete treatment of mathematical concept formation in young children, I would consider it necessary to probe more deeply into the issues of motivation, reinforcement and concept formation that surround the controversy between the discovery method and other more classical methods of reinforcement. Some experimental results on methods of immediate reinforcement are reported in the section on "Some Concept Experiments with Children."

I turn now to the specific topics I would like to develop more systematically. In the next section, a version of stimulus-sampling learning theory is formulated that holds considerable promise for providing a detailed analysis of the behavioral processes involved in the formation of mathematical concepts. In the following section, I report in somewhat abbreviated form six experiments dealing with mathematical concept formation in young children. A particular emphasis is placed on whether the learning process in this context is represented better by all-or-none or incremental conditioning. The final section is concerned with behavioral aspects of logical inference and, in particular, of mathematical proofs.

FUNDAMENTAL THEORY

The fundamental theory I shall apply in later sections is a variant of stimulus-sampling theory first formulated by Estes (1960). The axioms given here are very similar to those found in Suppes and Atkinson (1960). I shall not discuss the significance of the individual axioms at length because this has been done in print by a number of people. The axioms I may mention, however, are based on the following postulated sequence of events occurring on a given trial of an experiment: The organism begins the trial in a certain state of conditioning. Among the available stimuli a certain set is sampled. On the basis of the sampled stimuli and their conditioning connections to the possible responses, a response is made. After the response is made, a reinforcement occurs that may change the conditioning of the sampled stimuli. The organism then enters a new state of conditioning ready for the next trial. The following axioms (divided into conditioning, sampling, and response axioms) attempt to make the assumptions underlying such a process precise (they are given in verbal form but it is a routine matter to translate them into an exact mathematical formulation):

Conditioning Axioms

C1. *On every trial each stimulus element is conditioned to at most one response.*

C2. *If a stimulus element is sampled on a trial, it becomes conditioned with probability* c *to the response (if any) that is reinforced on that trial; if it is already conditioned to that response, it remains so.*

C3. *If no reinforcement occurs on a trial, there is no change in conditioning on that trial.*

C4. *Stimulus elements that are not sampled on a given trial do not change their conditioning on that trial.*

C5. *The probability* c *that a sampled stimulus element will be conditioned to a reinforced response is independent of the trial number and the preceding pattern of events.*

Sampling Axioms

S1. *Exactly one stimulus element is sampled on each trial.*

S2. *Given the set of stimulus elements available for sampling on a trial, the probability of sampling a given element is independent of the trial number and the preceding pattern of events.*

Response Axioms

R1. *If the sampled stimulus element is conditioned to a response, then that response is made.*

R2. *If the sampled stimulus element is unconditioned, then there is a probability* p_i *that response* i *will occur.*

R3. *The guessing probability* p_i *of response* i, *when the sampled stimulus element is not conditioned, is independent of the trial number and the preceding pattern of events.*

Although not stated in the axioms, it is assumed that there is a fixed number of responses and reinforcements and a fixed set of stimulus elements for any specific experimental situation.

Axioms C5, S2, and R3 are often not explicitly formulated by learning theorists, but for the strict derivation of quantitative results they are necessary to guarantee the appropriate Markov character of the sequence of state-of-conditioning random variables. Axioms of this character are often called *independence-of-path assumptions.*

The theory formulated by these axioms would be more general if Axiom S1 were replaced by the postulate that a fixed number of stimuli is sampled on each trial or that stimuli are sampled with independent probabilities, and if Axiom R1 were changed to read that the probability of response is the proportion of sampled stimulus elements conditioned to that response, granted that some conditioned elements are sampled. For the experiments to be discussed in the next section this is not an important generalization and will not be pursued here. (From the historical standpoint the generalizations just mentioned actually were essentially Estes' original ones.) Nowadays, they are referred to as the assumptions of the component model of stimulus sampling. Axiom S1 as formulated here is said to formulate the pattern model, and the interpretation is that the organism is sampling on a given trial the pattern of the entire stimulating situation, at least the relevant pattern, so to speak. This pattern model has turned out to be remarkably effective in providing a relatively good, detailed analysis of a variety of learning experiments ranging from rats in T-mazes to two-person interaction experiments.

There is one other general remark I would like to make before turning to the discussion of particular experiments. The kind of stimulus-response theory just formulated is often objected to by psychologists interested in cognitive processes. I do not doubt that empirical objections can be found to stimulus-response theory when stated in too simple a form. I am prepared, however, to defend the proposition that, at the present time, no other theory in psychology can explain in the same kind of quantitative detail an equal variety of learning experiments, including concepts formation experiments. I should also add that I do not count as different, cognitive formulations that are formally isomorphic to stimulus-sampling theory. In our recent book Atkinson and I (Suppes & Atkinson, 1960) attempted to show how the hypothesis language favored by many people (e.g., Bruner, Goodnow, & Austin, 1956) can be formulated in stimulus-sampling terms. For example, a strategy in the technical sense corresponds precisely to a state of conditioning and a hypothesis to the conditioned stimulus sampled on a given trial, but details of this comparison are not pertinent here.

SOME CONCEPT EXPERIMENTS WITH CHILDREN

I now turn to the applications of the fundamental theory, stated in the preceding section, to a number of experiments that are concerned with concept formation in young children. It would be possible, first, to describe these experiments without any reference to the theory, but, in order to provide a focus for the limited amount of data it is feasible to give in this survey, it will be more expedient to specialize the theory initially to the restricted one-element model, and report on data relevant to the validity of this model.

We obtain the one-element model by extending the axioms given in the preceding section in the following respect: we simply postulate that there is exactly one stimulus element available for sampling on each trial and that at the beginning of the experiment this single element is unconditioned.

This special one-element model has been applied with considerable success by Bower (1961) and others to paired-associate experiments, that is, to experiments in which the subject must learn an arbitrary association established by the experimenter between, say, a nonsense syllable as single stimulus and a response, such as one of the numerals 1–8 or the pressing of one of three keys. The most important psychological implication of this one-element model is that in the paired-associate situation the conditioning takes place on an all-or-none basis. This means that prior to conditioning the organism is simply guessing the correct response with the probability p_i mentioned in Axiom R3, and that the probability of conditioning on each trial in which the stimulus is presented is c. Once the stimulus is conditioned the correct response is made with probability one.

In an earlier paper Rose Ginsberg and I (Suppes & Ginsberg, 1963) analyzed a number of experiments, including some of those reported here, to exhibit a simple but fundamentally important fact about this all-or-none conditioning model. The assumptions of the model imply that the sequence of correct and incorrect responses prior to the last error form a binomial distribution of Bernoulli trials with parameter p. This null hypothesis of a fixed binomial distribution of responses prior to the last error admits, at once, the possibility of applying many powerful classical statistics that are not usually applicable to learning data. What is particularly important from a psychological standpoint is this hypothesis' implication that the mean learning curve, when estimated over responses prior to the last error, is a horizontal line. In other words, no effects of learning should be shown prior to conditioning. Ginsberg and I analyzed experiments concerned with children's concept formation, animal learning, and probability learning, and with paired-associate learning in adults from this standpoint. I shall not propose to give as extensive an analysis of data in the present paper as we attempted there, but I will attempt to cite some of the results on this question of stationarity because of its fundamental importance for any psychological evaluation of the kind of processes by which young children acquire concepts.

Other features of the experiments summarized below will be mentioned seriatim, particularly if they have some bearing on pedagogical questions. One general methodological point should be mentioned, however, before individual experiments are described. In many of the experiments, the stimulus displays were different on every trial so that there was no possibility of establishing a simple stimulus-response association. How is the one-element model to be applied to such data? The answer represents, I think, one of our more important general findings: *a very good account of much of the data may be obtained by treating the concept itself as the single element.* The schema, then, is that a simple concept-response association is established. With the single exception of Experiment I, we have applied this interpretation to the one-element model in our experiments.

Experiment I. Binary Numbers

This experiment is reported in detail in Suppes and Ginsberg (1962a). Five- and six-year-old subjects were required to learn the concepts of the numbers 4 and 5 in the binary number system, each concept being represented by three different stimuli; for example, if the stimuli had been chosen from the Roman alphabet, as in fact they were not, 4 could have been represented by abb, cdd, and eff, and 5 by aba, cdc, and efe. The child was required to respond by placing directly upon the stimulus one of two cards. On one card was inscribed a large Arabic numeral 4 and on the other a large Arabic numeral 5. All the children were told on each trial whether they made the correct or incorrect response, but half of them were also required to *correct* their wrong responses. Thus, in this experiment, in addition to testing the one-element model, we were concerned with examining the effect upon learning of requiring the subject to correct overtly a wrong response. There were 24 subjects in each of the two experimental groups. From test responses, after each experimental session, it seemed evident that whereas some subjects in both groups learned the concept as such, others learned only some of the specific stimuli representing the concepts so that, in effect, within each group there were two subgroups of subjects. It is interesting to note that this finding agrees with some similar results in lower organisms (Hull & Spence, 1938) but is contrary to results obtained with adult subjects for whom an overt correction response seems to have negligible behavioral effects (Burke, Estes, & Hellyer, 1954).

The data for both correction and non-correction groups are shown in Figure 1. It is apparent that there was a significant difference between the two groups in the rate of learning. The t of 4.00 computed between over-all responses of the two groups is significant at the .001 level.

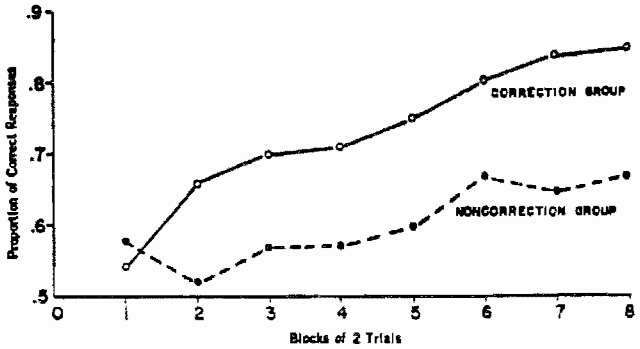

FIGURE 1.—Proportion of correct responses over all trials (binary-number experiment).

For the analysis of paired associates and concept formation we restricted ourselves to the 24 subjects of the correction group. To begin with, we analyzed the data as if each of the six stimuli, three for each number, represented an independent paired-associate item. In accordance with this point of view, we have shown in Figure 2 the proportion of correct responses prior to the last error and the mean learning curve for all responses.

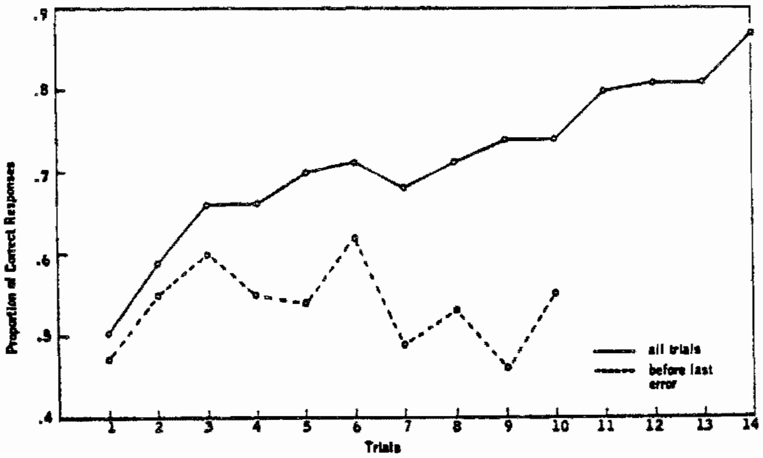

FIGURE 2.—Proportion of correct responses prior to last error and mean learning curve (binary-number experiment).

The data points are for individual trials. Because a total of only 16 trials were run on each stimulus we adopted a criterion of six successive correct responses, and thus the proportion of correct responses prior to the last error is shown only for the first 10 trials. A χ^2 test of stationarity over blocks of single trials supports the null hypothesis ($\chi^2 = 8.00$, $df = 9$, $P > .50, N = 844$).

Let us now turn to the question of concept formation. The identification we make has already been indicated. We treat the concept itself as the single stimulus, and in this case we regard the experiment as consisting of two concepts, one for the number 4 and one for the number 5. (It should be apparent that the identification in terms of the numbers 4 and 5 is not necessary; each concept can be viewed simply as an abstract pattern.)

The criterion for the learning of the concept was correct responses to the last three presentations of each stimuli. On this basis we divided the data into two parts. The data from the group meeting the criterion were arranged for concept-learning analysis—in this case a two-item learning task. The remaining data were assumed to represent paired-associate learning involving six independent stimulus items. For the paired-associate group over the first 10 trials we had 81 cases; for the concept-formation group we had 21 cases with 48 trials in each. The χ^2 test of stationarity was not significant for either group (for the concept subgroup $\chi^2 = 8.36$, $df = 9$, $P > .30$, $N = 357$; for the paired-associate subgroup $\chi^2 = 11.26$, $df = 8$, $P > .10$, $N = 570$).

To provide a more delicate analysis of this important question of stationarity we can construct Vincent curves in the following manner (cf. Suppes & Ginsberg, 1963). The proportion of correct responses prior to the last error may be tabulated for percentiles of trials instead of in terms of

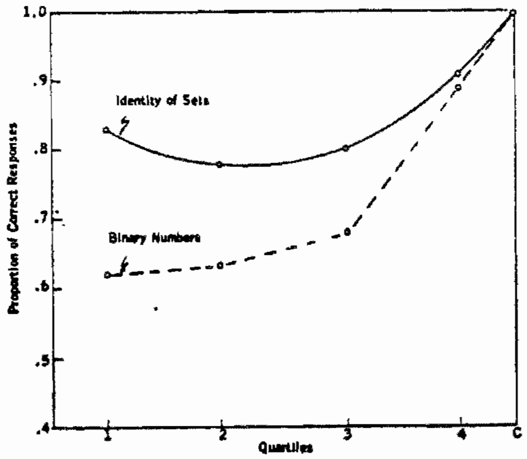

FIGURE 3.—Vincent learning curves in quartiles for proportion of correct responses prior to last error, binary numbers and identity of sets (Exps. I and II).

the usual blocks of trials. In Figure 3 the mean Vincent curve for the subjects in the binary-number experiment who met the concept criterion is shown. The curve is plotted in terms of quartiles. As the mean percentile of each of the four quartiles is 12.5 per cent, 37.5 per cent, 62.5 per cent, and 87.5 per cent, respectively, and C represents the 100 per cent point, the distance between 4, the fourth quartile, and C on the abscissa is one-half of that between the quartiles themselves. The evidence for nonstationarity in the final quartile will be discussed subsequently along with the other Vincent curve shown in this figure.

It should be noted, of course, that the subjects who take longer to meet the criterion are weighted more heavily in the Vincent curves. For example, suppose one subject has 16 responses prior to his last error whereas another subject has only 4. The first subject contributes 4 responses to each quartile whereas the second subject contributes only 1. This point will be discussed in more detail below. I turn now to the second experiment.

Experiment II. Equipollence and Identity of Sets

This experiment was performed with Rose Ginsberg and has been published in Suppes and Ginsberg (1963). The learning tasks involved in the experiment were equipollence of sets and the two related concepts of identity of sets and identity of ordered sets.

This subjects were 96 first-graders run in 4 groups of 24 each. In Group 1 the subjects were required to learn identity of sets for 56 trials and then equipollence for a further 56 trials. In Group 2 this order of presentation was reversed. In Group 3 the subjects learned first identity of ordered sets and then, identity of sets. In Group 4 identity of sets preceded identity of ordered sets. Following our findings in Experiment I, that is, that learning was more rapid when the child was required to make an overt correction response after an error, we included this requirement in Experiment II and most of the subsequent experiments reported below. Also, in this experiment and those reported below, no stimulus display on any trial was repeated for an individual subject. This was done in order to guarantee that the learning of the concept could not be explained by any simple principles of stimulus-response association, as was the case for Experiment I. For convenience of reference we termed concept experiments in which no stimulus display was repeated *pure* property of *pure* concept experiments.

The sets depicted by the stimulus displays consisted of one, two, or three elements. On each trial two of these sets were displayed. Minimal instructions were given the subjects to press one of two buttons when the stimulus pairs presented were "the same" and the alternative button when they were "not the same."

Our empirical aims in this experiment were several. First, we wanted to examine in detail if the learning of simple set concepts by children of this age took place on an all-or-none conditioning basis. Second, as the two sequences of learning trials on two different concepts for each group would indicate, we were interested in questions of transfer. Would the learning of one kind of concept facilitate the learning of another, and were there significant differences in the degree of this facilitation? Third, we were concerned with considering the question of finding the behavioral level at which the concepts could be most adequately defined. For example, in learning the identity of sets could the learning trials be satisfactorily analyzed from the standpoint of all trials falling under a single concept? Would it be better to separate the trials on which identical sets were presented from those on which nonidentical sets were presented in order to analyze

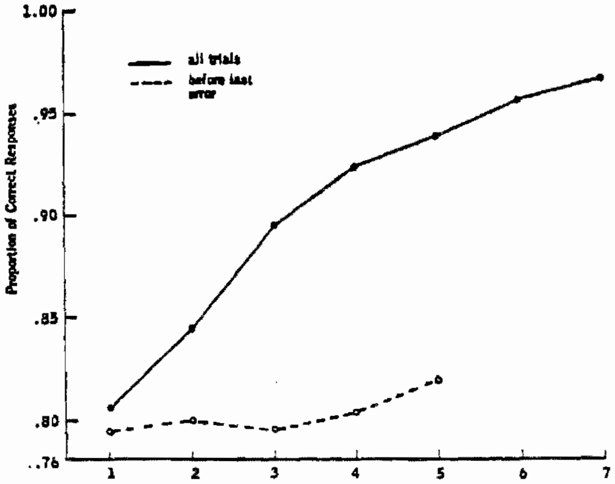

FIGURE 4.—Proportion of correct responses over all trials and before last error in blocks of eight trials, identity of sets, $N = 48$, Groups 1a and 4a (Exp. II).

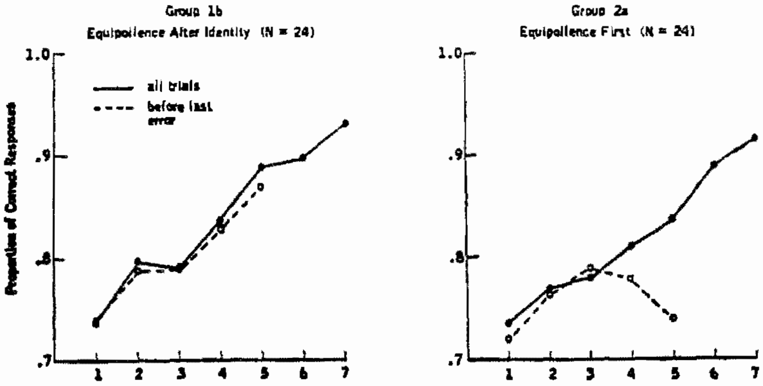

FIGURE 5.—Proportion of correct responses over all trials and before last error in blocks of eight trials, equipollence of sets, Groups 1b and 2a (Exp. II).

the data in terms of two concepts? Or would a still finer division of concepts in terms of sets identical in terms of order, sets identical as nonordered sets, equipollent sets and nonequipollent sets, be desirable?

In somewhat summary fashion the experimental results were as follows: The mean learning curves over all trials for all four groups are shown in Figures 4–7. As is evident from these curves the number of errors on the concept of identity of ordered sets was extremely small. From the high proportion of correct responses even in the first block of trials it is evident that this concept is a very natural and simple one for children. Learning curves for trials before the last error are also shown in these figures. To identify the last error prior to conditioning, we adopted a criterion of 16 successive correct responses. For this reason, these curves are only shown

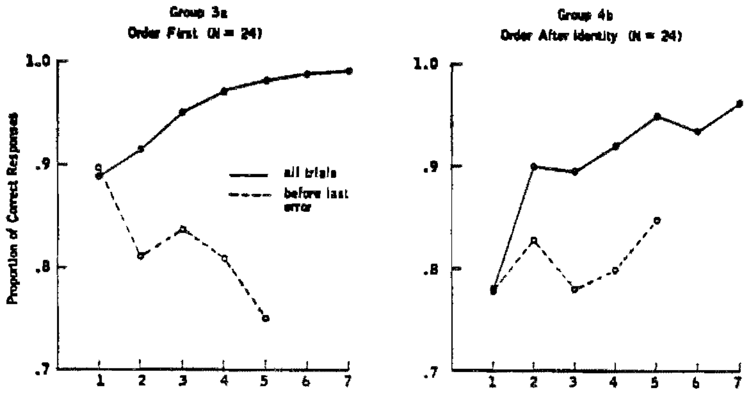

FIGURE 6.—Proportion of correct responses over all trials and before last error in blocks of eight trials, identity of ordered sets, Groups 3a and 4b (Exp. II).

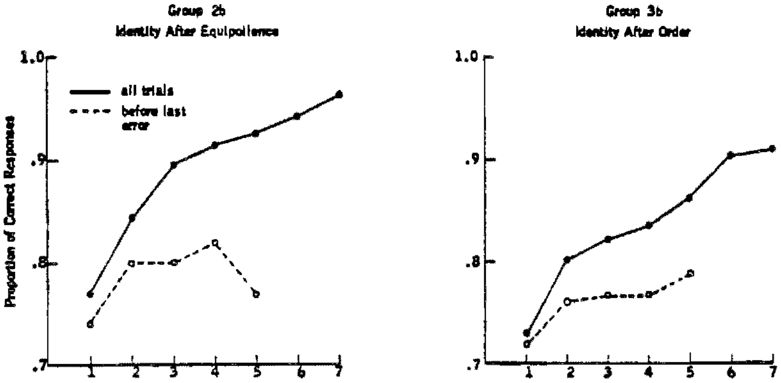

FIGURE 7.—Proportion of correct responses over all trials and before last error in blocks of eight trials, identity of sets, Groups 2b and 3b (Exp. II).

for the first 40 trials. The combined curve for Groups 1a and 4a is clearly stationary. This is also the case for 2b, 3a, 3b and 4b.[1] The results of the χ^2 test of stationarity for blocks of 4 trials are shown in Table 1 and confirm these graphic observations. Only the curve for 1b approaches significance. (No computation was made for 3a because of the small number of errors; the number of subjects in the final block of 4 trials is shown in the right-hand column of the table.)

TABLE 1

STATIONARITY RESULTS FOR EQUIPOLLENCE AND IDENTITY OF SETS EXPERIMENT (EXP. II)

Group	χ^2	*df*	$p>$	*S*s in last block
1a & 4a...	4.95	9	.80	9
1b...	16.69	9	.05	12
2a...	4.79	9	.80	11
3a...	—Too few errors—			1
4b...	4.89	9	.80	5
2b...	5.96	9	.70	5
3b...	3.49	9	.90	10

I shall restrict myself to one Vincent curve for this experiment. The 48 subjects of Groups 1 and 4 began with the concept of identity of sets. Of the 48 subjects, 38 met the criterion of 16 successive correct responses mentioned above. The Vincent curve for the criterion subjects is shown in Figure 3. Evidence of nonstationarity in the fourth quartile is present as in the case of Experiment I.

Examination of the mean learning curves over all trials apparently indicates little evidence of transfer. Somewhat surprisingly, the only definite evidence confirms the existence of negative transfer. In particular, it seems clear from Figure 6, there is negative transfer in learning the concept of identity of ordered sets after the concept of identity of unordered sets. Also, from Figures 4 and 7, it seems apparent that there is negative transfer in learning identity of sets after identity of ordered sets, but not after equipollence of sets.

The effects of transfer are actually more evident when we examine the data from the standpoint of two or four concepts. The mean learning curves over all 56 trials for the various concepts are shown in Figures 8-14. The data points are for blocks of 8 trials. The abbreviations used in the legends are nearly self-explanatory. For the learning curves shown at the right of each figure, the O curve is for pairs of sets identical in the sense of ordered sets, the IŌ curve for pairs of sets identical only in the sense of unordered sets, the EI curve for pairs of equipollent but not identical sets, and the Ē

[1] "Group 1a" refers to the performance of Group 1 subjects on the first of their two tasks, 1b to performance on the second task, and similarly for 2a, 2b, 3a, 3b, 4a, and 4b.

curve for pairs of nonequipollent sets. These four curves thus represent all pairs of sets in four mutually exclusive and exhaustive classes. The legend is the same for all figures. On the other hand, the curves for the two-concept analysis shown at the left of each figure differ in definition according to the problem being learned. In Figure 8 the dichotomy is

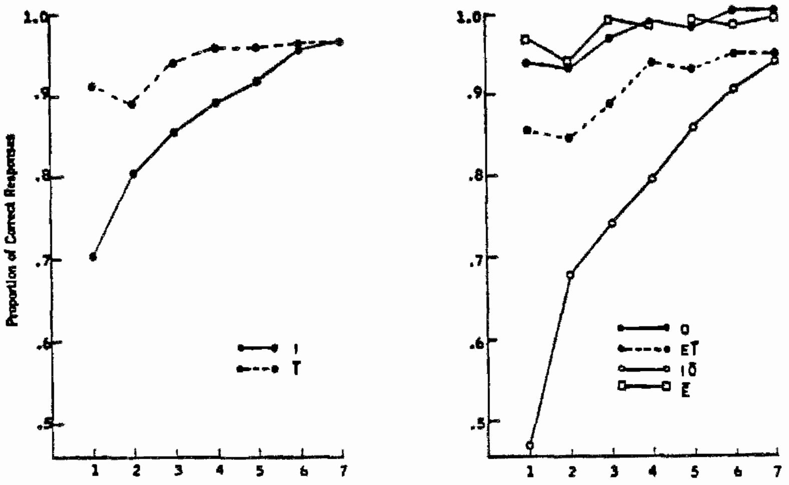

FIGURE 8.—Proportion of correct responses in blocks of eight trials for two and four concepts, identity of sets ($N = 48$), Groups 1a and 4a (Exp. II).

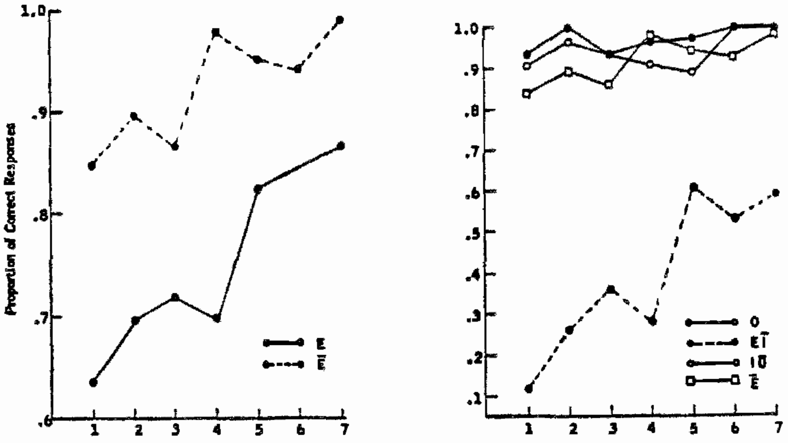

FIGURE 9.—Proportion of correct responses in blocks of eight trials for two and four concepts, equipollence following identity of sets, Group 1b (Exp. II).

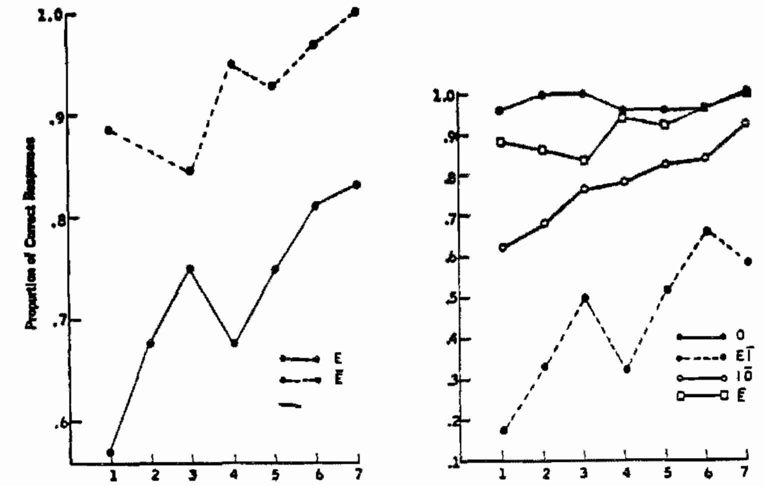

FIGURE 10.—Proportion of correct responses in blocks of eight trials for two and four concepts, equipollence, Group 2a (Exp. II).

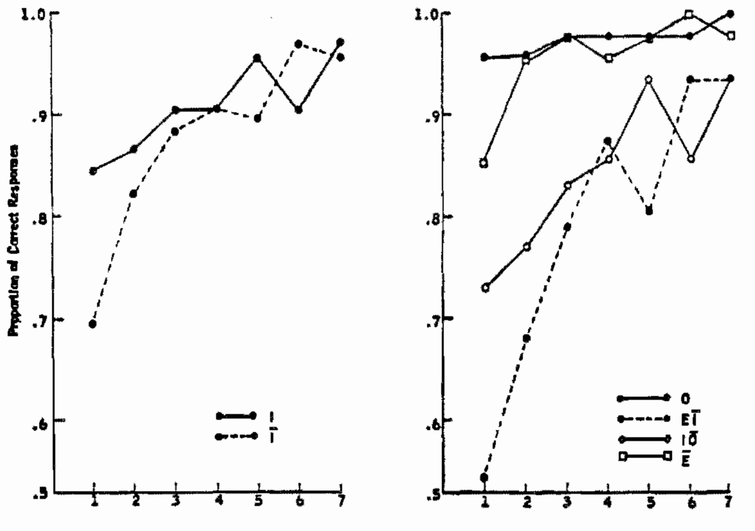

FIGURE 11.—Proportion of correct responses in blocks of eight trials for two and four concepts, identity following equipollence of sets, Group 2b (Exp. II).

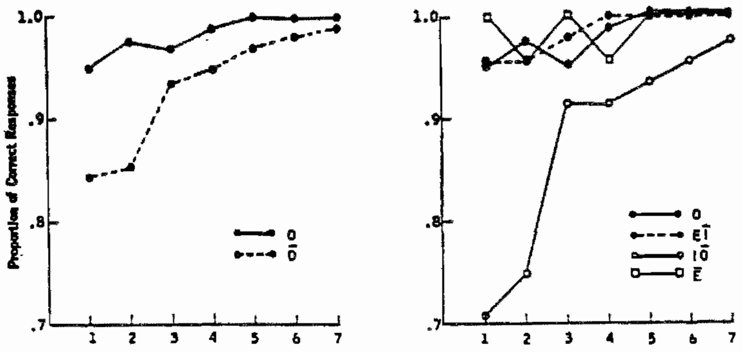

FIGURE 12.—Proportion of correct responses in blocks of eight trials for two and four concepts, identity of ordered sets, Group 3a (Exp. II).

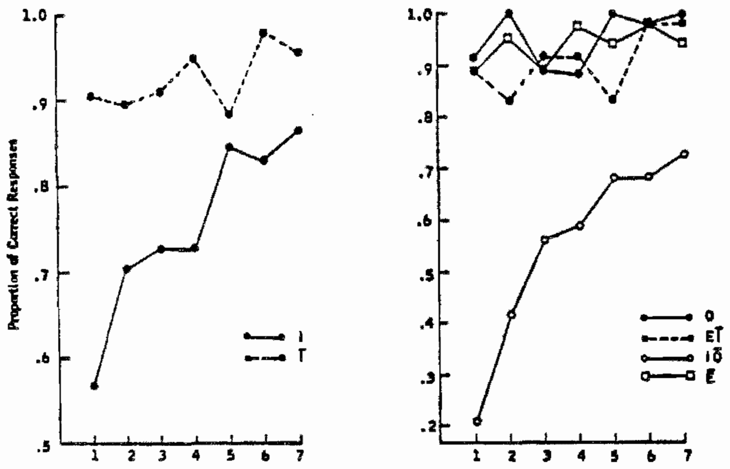

FIGURE 13.—Proportion of correct responses in blocks of eight trials for two and four concepts, identity of sets, following identity of ordered sets, Group 3b (Exp. II).

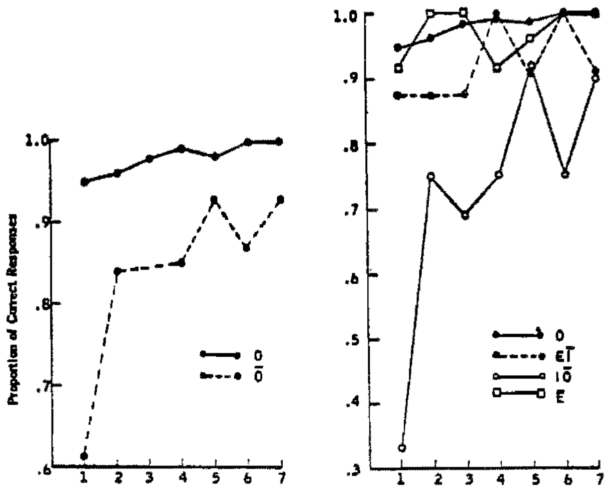

FIGURE 14.—Proportion of correct responses in blocks of eight trials for two and four concepts, identity of ordered sets following identity of sets, Group 4b (Exp. II).

identical and nonidentical sets (I and Ī); in Figure 9 it is equipollent and nonequipollent sets (E and Ē), and so forth for the other five figures.

Before considering questions of transfer, several observations should be made about the individual figures. First, for each of the eight subgroups (la–4b) the learning curves for the two-concepts and the four-concepts are not homogeneous. A difference in difficulty at either level of analysis can be detected in all cases. Second, contrary to some experimental results in concept formation, the two-concept curves at the left of each figure show that the absence of identity or equipollence is often easier to detect than its presence. The dichotomy of O vs. Ō, that is, identity or nonidentity of ordered sets, is the natural one. When the "presence" of a concept disagrees with this natural dichotomy, as it does in the case of identity and equipollence of sets, it is more difficult to detect than the absence of the concept. This conclusion is borne out by Figures 8 and 10 for the groups beginning with identity and equipollence, respectively, as well as for Group 3b (Fig. 13), that was trained on ordered sets before identity of sets. This same conclusion even holds fairly well for the second sessions after training on some other concept (Figs. 9, 11, 12). Figure 14, that compares O and Ō after training on identity of sets, indicates, I think, the tentative conclusion to be drawn. *Whether the absence or presence of a concept is more difficult to learn depends much more on the previous training and experience of a subject than on the concept itself.* When we compare Figure 12 with Figure 14 we see that even the difference between O and Ō in Figure 12 is influenced by the prior training or identity, for the difference is greater in Figure 14, and surely this is so because the IŌ cases have to be reversed in going from sets to ordered sets.

Third, examination of the four-concept curves reveals a natural gradient of difficulty. We may apply something rather like Coombs's (1950) unfolding technique to develop an ordinal generalization gradient. The natural or objective order of the classes of pairs of sets is O, IŌ, EI, Ē. For any of the three concepts of sameness studied in the experiment, we may, without disturbing this objective ordering, characterize the classes exhibiting presence of the concept and those exhibiting its absence by cutting the ordering into two pieces. On a given *side* of the *cut*, as I shall call it, the nearer a class is to the cut the more difficult it is. Consider, to begin with, Figure 8. The task is identity of sets, and the cut is between IŌ and EĪ; we see that, on the one side IŌ is more difficult than O, and on the other side of the cut, EĪ is more difficult than Ē. Turning to Figure 9, the task is equipollence and thus the cut is between EĪ and Ē; of the three concepts on the EĪ side, EĪ is clearly the most difficult and IŌ is slightly more difficult than O, sustaining the hypothesis of an ordinal gradient. In Figure 10, the task is equipollence again, but in this case without prior training, and the results are as expected but more decisive than those shown in Figure 9. Figure 11, like Figure 8, sustains the hypothesis when the task is identity of sets. In the case of Figure 12, the task is identity of ordered sets and thus IŌ, EĪ and Ē occur on the same side of the cut. IŌ is clearly the most difficult, but it is not really possible clearly to distinguish EĪ and Ē in difficulty, for very few errors are made in either class. In Figure 13 the task is identity of sets again, but this time following identity of ordered sets. The proper order of difficulty is maintained but the distinction between EĪ and Ē is not as sharply defined as in Figure 8 or Figure 11. Finally, in Figure 14, the task is identity of ordered sets following identity of sets. The gradients are as predicted by the hypothesis and are better defined than in Figure 12—no doubt because of the prior training on identity of sets. The existence and detailed nature of these natural gradients of difficulty within a concept task are subjects that seem to be worth considerable further investigation.

I turn now to evidence of transfer in the four-concept analysis. From examination of the over-all, mean learning curves which, in the terminology of the present discussion, are the one-concept curves, we observed no positive transfer but two cases of negative transfer. As might be expected, the four-concept curves yield a richer body of results. I shall try to summarize only what appear to be the most important points. Comparing Figures 8 and 11, we see that for the learning of identity of sets, prior training on equipollence has positive transfer for class IŌ and negative transfer for EĪ. The qualitative explanation appears obvious: the initial natural dichotomy seems to be O, Ō, and for this dichotomy IŌ is a class of "different" pairs, but the task of equipollence reinforces the treatment of IŌ pairs as the "same"; the situation is reversed for the class $E\bar{I}$, and thus the negative transfer, for under equipollence $E\bar{I}$ pairs are the "same," but under identity of sets they are "different."

Comparing now Figures 8 and 13 in which the task is again identity of sets but the prior training is on identity of ordered sets rather than equipollence, there is, as would be expected by the kind of argument just given, negative transfer for the class $I\bar{O}$. There is also some slight evidence of positive transfer for EI.

Looking next at Figures 9 and 10, we observe positive transfer for the class $I\bar{O}$ when the task is equipollence and the prior training is on identity of sets. What is surprising is the relatively slight amount of negative transfer for the class $E\bar{I}$.

Finally, we compare Figures 12 and 14, in which the task is identity of ordered sets; in the latter figure this task is preceded by identity of sets and we observe negative transfer for the class $I\bar{O}$, as would be expected. The response curves for the other three classes are too close to probability 1 to make additional inferences, although there is a slight negative transfer for $E\bar{I}$ that cannot be explained by the principles stated above.

It seems apparent from these results that the analysis of transfer in the learning of mathematical concepts may often be facilitated if a fine-scale breakdown of the concepts in question into a number of subconcepts is possible. Needed most is a quantitative theory to guide a more detailed analysis of the transfer phenomena.

Experiment III. Polygons and Angles

This experiment is reported in detail in Stoll (1962), and some of the data is presented here with her permission. The subjects were 32 kindergarten children divided into two equal groups. For both groups the experiment was a successive discrimination, three-response situation, with one group discriminating between triangles, quadrilaterals, and pentagons, and the other group discriminating between acute, right, and obtuse angles. For all subjects a typical case of each form (that is, one of the three types of polygons or three types of angles) was shown immediately above the appropriate response key. As in the case of Experiment II, no single stimulus display was repeated for any one subject. Stimulus displays representing each form were randomized over experimental trials in blocks of nine, with three of each type appearing in each block. The subjects were run to a criterion of nine successively correct responses, but with not more than 54 trials in any one session.

For the quadrilaterals and pentagons, the guessing probability prior to the last error was essentially the same, $\hat{p} = .609$ and $\hat{p} = .600$, respectively. Consequently, the proportions of correct responses for the combined data are presented in blocks of six trials, together with the mean learning curve for all trials, in Figure 15. The corresponding data for the triangles are not

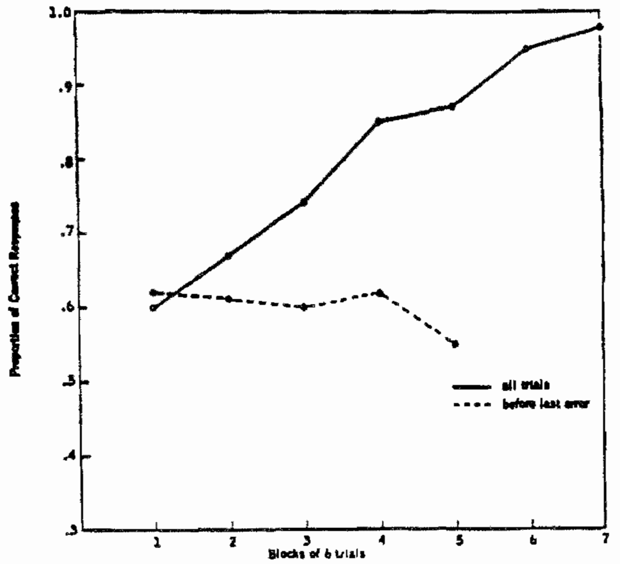

FIGURE 15.—Proportion of correct responses prior to last error and meaning learning curve (quadrilateral and pentagon concepts, Stoll experiment).

presented because the initial proportion of correct responses was quite high and the subjects learned to recognize triangles correctly very easily.

Fig. 16 presents the same curves for the combined data for the three types of angles, although the guessing probability varied between the

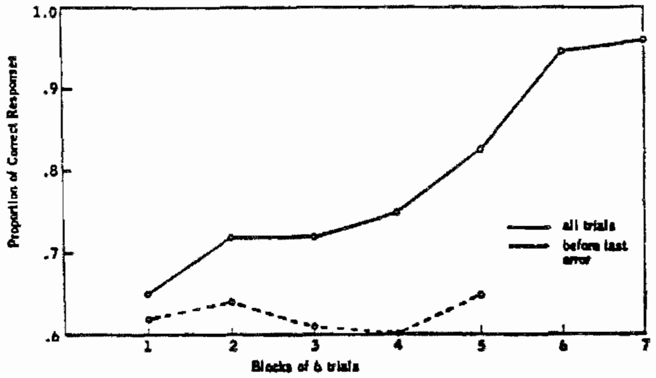

FIGURE 16.—Proportion of correct responses prior to last error and meaning learning curve (acute, right, and obtuse angle concepts, Stoll experiment).

angles. Both figures strongly support the hypothesis of a constant guessing probability prior to conditioning. In the case of the quadrilaterals and pentagons, $\chi^2 = 0.71$, $df = 4$, $P > .90$, $N = 548$. In the case of the combined data for the angles, $\chi^2 = 0.97$, $df = 4$, $P > .90$, $N = 919$.

The Vincent curves for each concept (except that of the triangle) are shown in Figure 17. The pentagons, quadrilaterals, and right angles have quite stationary Vincent curves, whereas there is a definite increase in the fourth quartile of the Vincent curves for the acute and obtuse angles, and in the case of the obtuse angles there is, in fact, a significant increase in the third quartile. Statistical tests of stationarity of these Vincent curves support the results of visual inspection. Each test has 3 degrees of freedom because the analysis is based on the data for the four quartiles. In the case of the quadrilaterals, $\chi^2 = 1.75$; for the pentagons, $\chi^2 = 1.33$; for the right angles, $\chi^2 = 0.95$; for the obtuse angles, $\chi^2 = 12.63$; and for the acute angles, $\chi^2 = 16.43$. Only the last two values are significant.

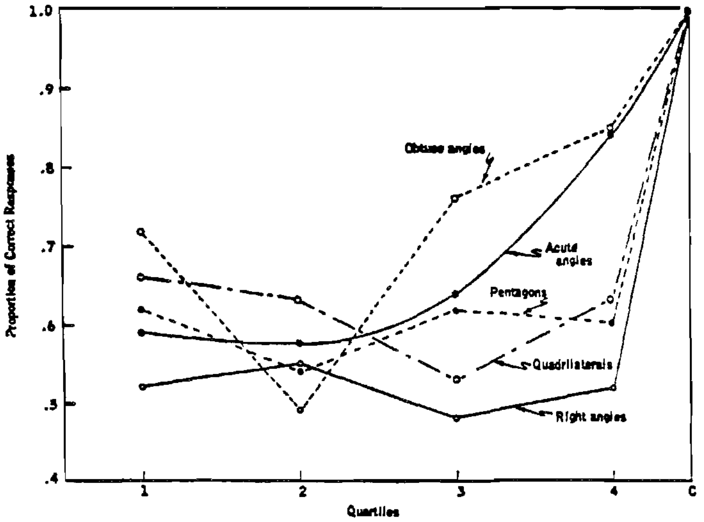

FIGURE 17.—Vincent learning curves in quartiles for proportion of correct responses prior to last error for Stoll experiment.

Using responses before the last error, for all concepts except that of triangle, goodness-of-fit tests were performed for (1) stationarity in blocks of six trials, (2) binomial distribution of responses as correct or incorrect in blocks of four trials, and (3) independence of responses, the test made for zero-order vs. first-order dependence. The results of these tests are presented in Table 2. The results shown strongly support the adequacy of the one-element model for this experiment.

Experiment IV. Variation in Method of Stimulus Display

In this study conducted with Rose Ginsberg, we compared the rate of learning in two experimental situations, one in which stimulus displays were

TABLE 2

STATIONARITY, ORDER, AND BINOMIAL DISTRIBUTION RESULTS (STOLL EXPERIMENT ON GEOMETRIC FORMS)

	X^2	*df*	$P >$
Quadrilateral, p = .609:			
Stationarity (N = 273)	1.68	4	.70
Order (N = 262)	0.65	1	.40
Binomial distribution (N = 65)	1.77	2	.40
Pentagon, p = .600:			
Stationarity (N = 275)	2.40	4	.60
Order (N = 269)	1.76	1	.15
Binomial distribution (N = 65)	2.07	2	.35
Acute angle, p = .674:			
Stationarity (N = 338)	7.96	4	.05
Order (N = 348)	3.17	1	.05
Binomial distribution (N = 85)	2.66	2	.25
Right angle, p = .506:			
Stationarity (N = 313)	6.34	4	.10
Order (N = 326)	2.41	1	.10
Binomial distribution (N = 80)	10.52	2	.001*
Obtuse angle, p = .721:			
Stationarity (N = 268)	1.10	4	.85
Order (N = 256)	7.32	1	.001*
Binomial distribution (N = 63)	2.90	2	.20
Quadrilateral and pentagon, p = .604:			
Stationarity (N = 548)	0.71	4	.90
Binomial distribution (N = 130)	1.77	2	.40
All angles, p = .624:			
Stationarity (N = 919)	0.97	4	.90

presented individually in the usual way, and the other in which the same stimulus displays were presented by means of colored slides, to groups of four children. The concept to be learned was identity of sets, and in both situations the children were required to respond by pressing one of two buttons, depending upon whether the stimulus display on that trial was identical or non-identical. Of the 64 subjects 32 were from first grade and 32 from kindergarten classes. For the children receiving individual displays the experimental situation was essentially identical with that of Experiment II.

Each group, however, was divided into two subgroups. One subgroup received the stimulus material in random order, and the other in an order based on anticipated difficulty; in particular, presentations of one-element sets came first, then two-element sets, and finally three-element sets.

The mean learning curves for the two subgroups with random presentation are shown in Figure 18. The results suggest that presentation by slides is a less effective learning device for younger children, and the younger the child, the more this finding seems to apply. At all levels of difficulty, the kindergarten children learned more efficiently when the stimuli were presented to them in individual sessions. With one- or two-

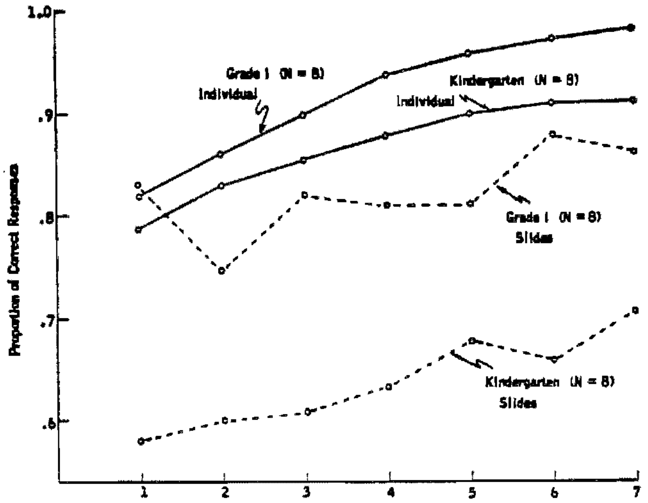

FIGURE 18.—Proportion of correct responses in blocks of 12 trials, subgroups with random presentation (Exp. IV).

element sets displayed, grade 1 subjects learned only slightly better in the individual session situation than in the slide situation, but when the task was more difficult (stimulus displays of three-element sets) the individual learning situation was clearly the most adequate. In interpreting these results it should be emphasized that the individual session was strictly experimental so that the amount of interaction between subject and experimenter was paralleled in both individual and slide situations.

Why these two experimental situations should produce different results in terms of learning efficiency is not yet clear to us. One possibility is the following: It has been shown, both with lower organisms (Murphy & Miller, 1955) and young children (Murphy & Miller, 1959), that the ideal situation for learning is the contiguity of stimulus, response and reinforcement. In the individual sessions these requirements were met, for the response buttons were 1.5 inches below the stimulus displays and the reinforcement lights were 1.0 inches from the stimuli. On the other hand, in the slide presentations, although the stimulus displays and reinforcements were immediately adjacent to each other, the response buttons were about 3 feet from the screen on which the stimulus display was projected. Experimentally, it has been shown (Murphy & Miller, 1959) that with children of this age group a separation of 6.0 inches is sufficient to interfere with efficient learning.

Experiment V. Incidental Learning

This experiment represents a joint study with Rose Ginsberg. Thirty-six kindergarten children, in 3 groups of 12 each, were run for 60 trials a day on 2 successive days of individual experimental sessions during which they were required to learn equipollence of sets. On the first day, the stimulus displays presented to the subjects on each trial differed in color among the three groups but otherwise were the same. In Group 1, all displays were in one color—black—and in Group 2, equipollent sets were red and nonequipollent sets, yellow. For the first 12 trials in Group 3, equipollent sets were red and nonequipollent sets, yellow; for the remaining 48 trials on that day the two colors were gradually fused until discrimination between them was not possible. On the second day, all sets were presented to all three groups in one color—black.

As is apparent from the brief description of the experimental design, Group 1 simply had two days' practice under the same conditions with the concept of equipollence. In Group 2, the child did not actually need to learn the concept of equipollence but could simply respond to the color difference on the first day. It is well known that such a color discrimination for young children is a simple task. If the child in this group learned anything about equipollence of sets the first day, therefore, we may assume it to have been a function of incidental learning. If incidental learning is effective, his performance on the second day, when the color cue is dropped, should have been at least better than the performance of children in Group 1 on the first day. In Group 3, where we gave the child the discriminative cue of color difference in the first trial and then very slowly withdrew that cue, the child should have continued to search the stimulus displays very closely for a color stimulus and thus have been obliged to pay close attention to the stimuli.

The mean learning curve for the three groups are shown in Figure 19.

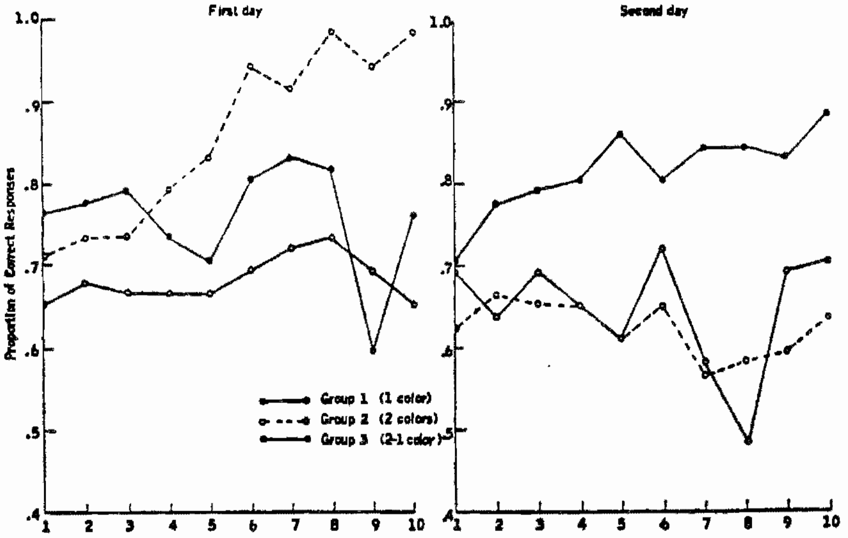

FIGURE 19.—Proportion of correct responses in blocks of six trials for both days (Exp. V).

Of the three groups only Group 2 approached perfect learning on the first day. In this group, of course, only color discrimination was necessary. Both the other groups did not improve over the first 60 trials, although Group 3 showed some initial improvement when the color cues remained discriminable. On the second day, Group 1 showed no improvement, and the learning curves for this group and Group 2 were practically identical. For Group 3, on the other hand, the results were conspicuously better on the second day than for those of any other group. It is apparent from these curves that the task chosen was relatively difficult for the age of the children because essentially no improvement was shown by Group 1 over the entire 120 trials. The conditions in Group 3, where the children were forced to pay very close attention to the stimuli, do seem to have significantly enhanced the learning.

Experiment VI. Variation of Response Methods

This study was made jointly with Rose Ginsberg. Its object was to study the behavioral effects of different methods of response. Specifically, 3 groups, each composed of 20 kindergarten children, were taken individually through a sequence of 60 trials on each of 2 successive days for a total of 120 trials. The task for all 3 groups was equipollence of sets.

In Group 1, the child was presented with pictures of two sets of objects and was to indicate, by pressing one of two buttons, whether the sets "went together" or did not "go together" (were equipollent or nonequipollent).

In Group 2, the child was presented with one display set and two "answer" sets and was required to choose the answer that "went together" with the display set.

In Group 3, the child was presented with one display set and three "answer" sets and was to make his choice from the three possible answers.

This situation has fairly direct reference to teaching methodology in the sense that Group 2 and Group 3 represent multiple-choice possibilities. In Group 1, where the child is required to identify either the presence of the concept or its absence on each trial, the situation is comparable to one in which the child must indicate whether an equation or statement is correct or incorrect.

On the first day, each group of children learned the task described above. On the second day, they were run on an alternative method. Specifically, Group 1 was run under Group 3 conditions and Groups 2 and 3 were run under Group 1 conditions.

The mean learning curves for all groups on both days are shown in Figure 20. It will be noticed that in Group 2, where the subjects were required to choose from one of two available responses, they learned slightly more quickly and to a slightly better level of achievement on the first day than the other groups but, on the second day, when the experimental conditions were shifted, Group 2 subjects did less well than the subjects in the other two groups. The clear superiority of Group 1 on the second day, when

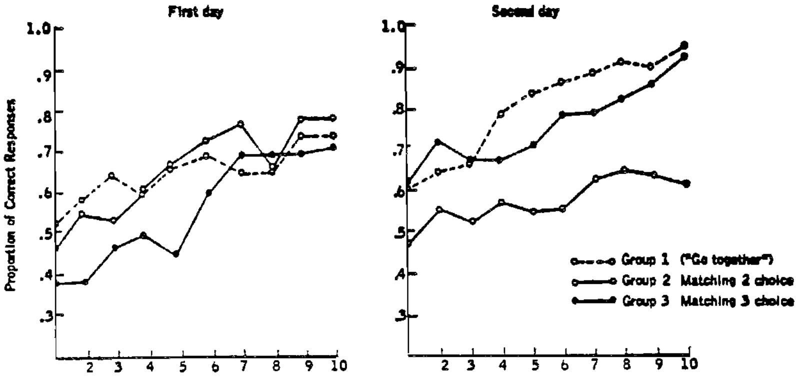

FIGURE 20.—Proportion of correct responses for two successive days in blocks of six trials for all subjects (Exp. VI).

they were transferred to Group 3 conditions, indicates some positive transfer from learning to judge whether or not a concept is present to the multiple-choice situation, whereas the results for Groups 2 and 3 on the second day indicate some negative transfer from the multiple-response methods to the presence-or-absence method.

These results are further supported when we examine separately the data from subjects achieving a criterion of 12 successive correct responses on the first day. The more successful method was clearly that used in Group I, as indicated by the curves in Figure 21. The subjects in this group were

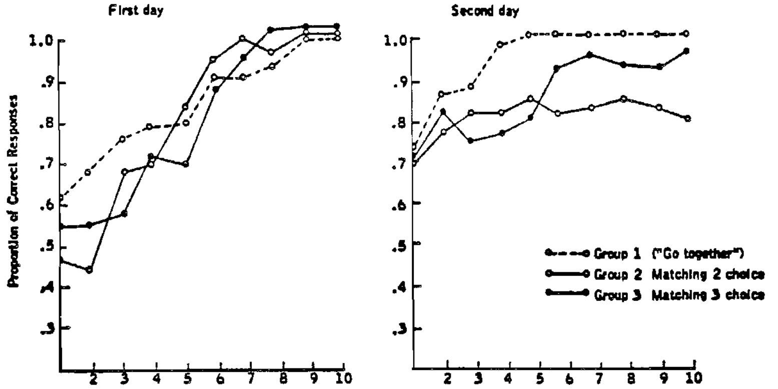

FIGURE 21.—Proportion of correct responses for two successive days in blocks of six trials for subjects achieving of 12 successive correct responses (Exp. VI).

conspicuously more successful than the other groups on the second day, making, in fact, no errors from Trial 30 to Trial 60. Group 3 achieved perfect scores on the second day only on the last six trials, and Group 2 never reached that level on the second day, although, like the other criterion subjects, they had achieved perfect learning on the first day.

It seems reasonable to conclude tentatively that the method used with Group 1, where subjects were required to recognize the presence or absence of some property on each trial, is the more successful method in establishing the understanding of a concept well enough to permit transfer to a different response method.

Support for the all-or-none model of conditioning is also to be found in this experiment. In Table 3, χ^2 goodness-of-fit tests of stationarity over trials before the final error for each group on each day are shown. The six values are all nonsignificant and thus support the basic assumption of the all-or-none models.

TABLE 3
TEST FOR STATIONARITY OVER TRIALS BEFORE THE FINAL ERROR (EXP. VI)

	Group 1		Group 2		Group 3	
	Day 1	Day 2	Day 1	Day 2	Day 1	Day 2
χ^2	4.97	2.41	10.76	4.255	16.07	2.87
df	8	1	9	8	9	7
$P >$	.70	.10	.20	.80	.05	.80

Some Tentative Conclusions

On the basis of the six experiments just discussed I would like to draw some tentative conclusions, some of which are important for pedagogical procedures (cf. Suppes and Ginsberg, 1962b). I want to emphasize, however, that I do not wish to claim that the evidence from these experiments is conclusive enough to establish any one of the six conclusions in any final way, but what I do hope is that the attempt to summarize some of the implications of these experiments will stimulate other research workers to investigate these and related propositions in more adequate detail.

1. Formation of simple mathematical concepts in young children is approximately an all-or-none process. Evidence indicates, however, that significant deviations from the all-or-none model are present (see the discussion of the two-element model below).
2. Learning is more efficient if the child who makes an error is required to make an overt correction response in the presence of the stimulus to be learned (Exp. I).
3. Incidental learning does not appear to be an effective method of acquisition for young children. In Experiment IV the group of children that responded to a color discrimination did not subsequently give any indication of having learned the underlying concepts.
4. Contiguity of response, stimulus, and reinforcement enhances learning (Exp. V).
5. In the learning of related mathematical concepts the amount of over-all transfer from the learning of one concept to another is surprisingly small. However, considerable positive or negative transfer between specific subconcepts is often present (Exp. II).

6. Transfer of a concept is more effective if, in the learning situation, the subject is required to recognize the presence or absence of a concept in a number of stimulus displays, than if learning has involved matching from a number of possible responses (Exp. VI).

Several of these conclusions are at variance with generally accepted results for adult learning behavior. For example, the efficacy of an immediate overt correction response (see Burke et al., 1954, for negative results on this method in adults), the variation of response method, or the relative specificity of the learning of concepts with relatively little transfer. What is much needed is a wider range of systematic studies to isolate the factors of learning in young children which are particularly distinct from common features of adult learning behavior.

Two-Element Model

In the first conclusion mentioned above, we stated that the formation of concepts is approximately an all-or-none process in young children. On the other hand, the detailed analysis of responses prior to the last error indicates that, in many cases, there is an incremental effect appearing in the last quartile or even, sometimes, in the last two quartiles of the data. This matter is discussed in some detail in Suppes & Ginsberg (1963). I would simply like briefly to mention here what currently appears to be the best extension of the one-element model to account for these results.

The simplest alternative model is the linear incremental model with a single operator. The intuitive idea of this model is precisely the opposite of the all-or-none conditioning model. The supposition is that learning proceeds on an incremental basis. Let q_n be the probability of an error on trial n. Then the model is formulated by the following recursive equation:

$$q_{n+1} = (1\text{-}\theta)q_n, \qquad (1)$$

where $O < \Theta \leq 1$. It is simple to show but somewhat surprising that this purely incremental model has precisely the same mean learning curve as the all-or-none model if we set $c = \Theta$. (To obtain this identity of the learning curves we must, of course, consider all responses and not simply responses prior to the last error.) The incremental model differs sharply from the all-or-none model in the kind of learning curve predicted for responses prior to the last error, as is evident from equation (1). It may be shown, moreover, that the concave upward Vincent curves obtained in several of the experiments discussed above (see Figs. 3 and 17) cannot be accounted for by the linear incremental models.

The second simple alternative, that will account for these concave-upward Vincent curves, is a model that represents a kind of compromise between the all-or-none model and the incremental model. It results from a simple extension of the one-element model, that is, the assumption that associated with each situation are two stimulus elements and, therefore, learning proceeds in two stages of all-or-none conditioning. Each of the two elements is conditioned on an all-or-none basis but the two parameters of conditioning, one for each element, may be adjusted to produce various incremental effects on the response probabilities. Let σ and τ be the two elements. The basic learning process may be represented by the following four-state Markov process in which the four states (σ,τ), σ, τ, and O represent

$$\begin{array}{c|cccc} & (\sigma,\tau) & \sigma & \tau & 0 \\ \hline (\sigma,\tau) & 1 & 0 & 0 & 0 \\ \sigma & b'/2 & 1-b'/2 & 0 & 0 \\ \tau & b'/2 & 0 & 1-b'/2 & 0 \\ 0 & 0 & a/2 & a/2 & 1-a \end{array}$$

the possible states of conditioning of the two- stimulus elements. Because we do not attempt experimentally to identify the stimuli σ and τ, this Markov process may be collapsed into a three-state process, in which the states are simply the *number* of stimuli conditioned to the correct response. In the matrix shown above a is the probability of conditioning at the first stage and b′ is the probability of conditioning at the second stage. The division by ½ in the matrix simply represents the equal probability of sampling one of the two elements. If we consider only the number of stimuli, it is convenient to replace $b'/2$ by b and we obtain the transition matrix shown below:

$$\begin{array}{c|ccc} & 2 & 1 & 0 \\ \hline 2 & 1 & 0 & 0 \\ 1 & b & 1-b & 0 \\ 0 & 0 & a & 1-a \end{array}$$

To complete the description of the process we associate with the sampling of each element σ and τ a guessing probability g_σ and g_τ when the elements are still unconditioned. For the states 0 and 1 of the second matrix shown we then have the guessing probabilities g_0 and g_1 defined in the obvious manner in terms of the sampling probabilities:

$$g_0 = \tfrac{1}{2}\, g_\sigma + \tfrac{1}{2}\, g_\tau,$$

$$g_1 = \tfrac{1}{4}\, g_\sigma + \tfrac{1}{4}\, g_\tau + \tfrac{1}{2} = \tfrac{1}{2}\, g_0 + \tfrac{1}{2}.$$

The probabilities g_σ and g_τ are not observable but g_0 is, and g_1 is a simple function of it. This means that we now have a process with three free parameters, the conditioning parameters a and b and the guessing probability g_0. I shall not attempt to report on the detailed application of this two-element model, but we are now in the process of applying it to a number of different experimental situations and hope to report in detail on its empirical validity in the near future. I would, however, like to remark that a very interesting interpretation of this kind of two-stage model has recently been given by Restle (1964), who interprets the two stages of learning as conditioning and discrimination. The model he proposes differs in detail from that given here, but for most observable response patterns the differences between the two will not be large.

Before turning to another topic, I would like to emphasize that I do not feel that the analysis of concept formation in terms of the simple one- and two-element models sketched here is fully satisfactory intellectually. It is apparent that these models must be regarded as schemata of the full process that is taking place in concept formation. What is surprising is that they are able to account for response data as well as they do. Theories that postulate more details about the learning process in concept formation are needed to go beyond the present analysis. This, I take it, will be particularly true as we proceed to the analysis of more complicated mathematical concepts, whose learning must rest upon the understanding of simpler concepts.

LOGIC AND MATHEMATICAL PROOFS

Together with several younger associates I have conducted, for several years, pedagogical and psychological experiments on the learning of mathematical logic with elementary-school children. Before turning to a relatively systematic statement of some of our results, I would like to survey briefly what we have attempted.

In the fall of 1956 I brought into my college logic course a selected group of sixth-, seventh-, and eighth-graders (they were, in fact, no more selected than the Stanford students in the course). Their demonstrated ability to master the course and perform at a level only slightly below that of the college students was the initial impetus for further work. The next important step was the extensive study by Shirley Hill of the reasoning abilities of first-, second-, and third-graders. This study was begun in 1959 and completed as her dissertation in 1961. I shall report briefly on this below. In 1960 Dr. Hill and I wrote a text and taught a pilot group of fifth-graders a year's course in mathematical logic. The course was structured very similarly to a college logic course except that material was presented more explicitly and at a much slower pace. Students were selected on the basis of ability and interest (the minimum IQ was 110), and again the positive results were an impetus to further work. Because of the success of this class, the text book was revised (Suppes & Hill, 1964) and, during the academic year (1962–63), was taught to approximately 300 selected fifth-graders in the Bay Area, with support for the project coming from the Office of Education and the National Science Foundation. These same classes were given a second year of instruction as sixth-graders and, in another year, we shall be able to report in detail on their level of achievement. We were also interested in seeing if we could train fifth-grade teachers to teach the course as part of their regular curriculum. To this end, we gave them a special course in logic in the summer of 1961 and all the classes but one were taught by the teachers.

We began experimental psychological studies of how and to what degree children of still younger ages could learn the concepts of formal inference. I shall report briefly on a pilot study with first-graders. On the basis of the experience of several of us with the teaching of logic to elementary-school children, we conducted an extensive psychological experiment with fourth-grade children to determine whether it was easier initially to learn rules of sentential inference when the standard interpretations were given, or whether it was easier simply to learn the rules as part of an uninterpreted meaningless game. This last possibility was, of course, most disturbing for a wide variety of mathematicians interested in the teaching of mathematics. I shall not enter here into the many reasons why I think there are good psychological arguments to believe that the initial teaching of inference simply as a game will turn out to be the most effective approach. I am frankly reluctant to formulate any very definite ideas about this highly controversial matter until we have accumulated a much more substantial body of evidence.

I turn now to the two experiments mentioned above on which I want to report briefly.

Experiment VII. Logical Abilities of Young Children

As already remarked, this extensive empirical study constituted Shirley Hill's doctoral dissertation (1961). Dr. Hill gave a test instrument consisting of 100 items to 270 children in the age group 6 through 8 years (first, second, and third grades). Each of the 100 items consisted of 2 or 3 verbal premises plus a conclusion presented orally as a question. The subject was asked to affirm or deny the conclusion as presented. There were two primary reasons for not asking the children to compose a conclusion: In the first place, children of this age sometimes have difficulty formulating sentences; this has sometimes been cited as the reason for inappropriate measures of their reasoning abilities. The second reason is, simply, the methodological difficulty of interpreting the correctness or incorrectness of a conclusion given as a free response. The 100 items were equally divided between positive and negative answers. The first part of the test consisted of 60 items that were drawn from sentential logic. Every conclusion or its negation followed from the given premises by the sentential theory of inference. The second part consisted of 40 items that were drawn from predicate logic, including 13 classical syllogisms. The predicate logic items, however, also included inferences using two-place predicates together with existential quantifiers.

Because it is easy for children to give the correct answer to a problem in which the conclusion is generally true or false, every attempt was made to construct the items in such a way that the omission of one premise would make it impossible to draw the correct conclusion. To provide a behavioral check on this aspect of the items a base-line group of 50 subjects was given the test with the first premise of each item omitted. For instance, to quote the illustration given by Dr. Hill (1961, p. 43), the original item might read:

> If that boy is John's brother, then he is ten years old.
> That boy is not ten years old.
> Is he John's brother?

For the base-line group the item would be presented:

> If that boy is not ten years old, is he John's brother?

An example of a badly constructed item would be the following:

> If boys are stronger than girls, then boys can run faster than girls.
> Boys are stronger than girls.
> Can boys run faster than girls?

Naturally almost all children gave the correct answer to this latter item, but their behavioral response actually told us little about their intuitive grasp of principles of logical inference. That Dr. Hill's items were well constructed are attested to by the fact that the base-line group averaged 52.02 per cent correct items, which does not significantly differ from chance. (Note that this percentage is based on 5,000 subject items.)

I shall not go into all the facets of Dr. Hill's study here. I mainly want to report on one or two of the most important conclusions. Let me first mention the results of the three standard groups of ages 6, 7, and 8 years. The 6-year-old group receiving the items described above got 71.18 per cent of the items correct. The 7-year-old group got 79.54 per cent of the items correct, and the 8-year old group got 85.58 per cent correct. These percentage figures indicate a steady increase with age in the ability to draw correct logical inferences from hypothetical premises. In addition to the fact of increase, it is just as important to note that the 6-year-old children performed at quite a high level, in contradiction to the view of Piaget and his followers that such young children are limited to concrete operations. Dr. Hill's study certainly provides substantial evidence to the contrary.

To avoid any possible confusion, it should be borne in mind that no claim is made that this study shows young children to be able explicitly to state formal principles of inference. What is claimed is that their grasp of the structure of ordinary language is sufficiently deep for them to be able to make *use* of standard principles of inference with considerable accuracy.

I would like to present just two other results of Dr. Hill's study. To avoid the conjecture that children aged six may be able to do the simpler forms of inference quite well but will do badly on the more difficult inferences involving two-place predicates, the percentage of correct responses for each age group on the 10 types of inferences appearing in the 100-item test are shown in Table 4. The last two categories entitled "Quantificational Logic—Universal Quantifiers" and "Quantificational Logic—Existential Quantifiers" refer to inferences that do not fall within the scheme of the classical syllogism. Although these last two categories are more difficult

TABLE 4

Percentage of Correct Responses for Different Principles of Inference by Age Level

	Percentage of Correct Responses		
Principles of Inference	Age 6	Age 7	Age 8
Modus ponendo ponens	78	89	92
Modus tollendo ponens	82	84	90
Modus tollendo tollens	74	79	84
Law of hypothetical syllogism	78	86	88
Hypothetical syllogism and *tollendo tollens*	76	79	85
Tollendo tollens and *tollendo ponens*	65	77	81
Ponendo ponens and *tollendo tollens*	65	67	76
Classical syllogism	66	75	86
Quantificational logic—universal quantifiers	69	81	84
Quantificational logic—existential quantifiers	64	79	88

than the simplest *modus ponendo ponens* applications, the performance level of the children aged six is still well above chance, and it is interesting to note that the performance on universal quantifiers is actually slightly better than the performance on sentential inferences using both *ponendo ponens* and *tollendo ponens.*

The second result concerns the attempt to identify some of the more obvious sources of difficulty. The lack of a sharply defined gradient in Table 4 suggested further examination of individual items. What turned out to be a major source of difficulty was the inclusion of an additional negation in an inference. Two hypothetical items that illustrate this difference are the following: Consider first as a case of *modus ponendo ponens:*

If this is Room 7, then it is a first-grade room.
This is Room 7.
Is it a first-grade room?

Let us now modify this example, still making it an application of *modus ponendo ponens:*

If this is not Room 8, then it is not a first-grade room.
This is not Room 8.
Is it a first-grade room?

The additional negations in the second item are a source of considerable difficulty to the children. It might be thought that the negations simply cause difficulty because they represent an increase in general complexity. To examine this question Dr. Hill compared the cases using a single rule of inference in which negations occurred, with the use of combined implications involving more than one rule of inference. The results are shown in Table 5. It is clear from this table that an additional negation adds a greater factor of difficulty than the use of more than one principle of inference.

I have only presented here a few of the results of this important study. A complete statement of the results are included in Hill (1961).

TABLE 5

COMPARISON OF INCREASE IN ERROR ASSOCIATED WITH THE ADDITION OF NEGATION AND WITH COMPOUND IMPLICATIONS

	PERCENTAGE OF ERROR OUT OF TOTAL POSSIBLE RESPONSES		
PRINCIPLES OF INFERENCE	Regular Form	Additional Negation	Combined Implication
Modus ponendo ponens	.06	.19	.17
Modus tollendo tollens	.12	.34	
Modus tollendo ponens	.03	.25	.27
Modus tollendo tollens	.12	.34	
Law of hypothetical syllogism	.08	.22	.16
Modus tollendo tollens	.12	.34	

Experiment VIII. Pilot Study of Mathematical Proofs

The details of this pilot study are in Suppes (1962). The original study was conducted with the assistance of John M. Vickers, and we are now engaged in a larger study along the same lines. The primary objective of this pilot study was to determine if it is feasible to apply the one-element model, described earlier, to the behavior of young children by constructing proofs in the trivial mathematical system, described as follows: Any finite string of 1's is a well-formed formula of the system. The single axiom is the single symbol 1. The four rules of inference are:

$$\begin{array}{ll} \text{R1.} & S \rightarrow S11 \\ \text{R2.} & S \rightarrow S00 \\ \text{R3.} & S1 \rightarrow S \\ \text{R4.} & S0 \rightarrow S \end{array}$$

where S is a non-empty string. A theorem of the system is, of course, either the axiom or a finite string that may be obtained from the axiom by a finite number of applications of the rules of inference. A general characterization of all theorems is immediate: any finite string is a theorem if and only if it begins with 1. A typical theorem in the system is the following one, which I have chosen because it uses all four rules of inference:

Theorem	101
(1) 1	Axiom
(2) 100	R2
(3) 10	R4
(4) 1011	R1
(5) 101	R3

The proofs of minimal length in this system are easily found, and the correction procedure was always in terms of a proof of minimal length.

The stimulus discrimination facing the subject on each trial is simply described. He must compare the last line of proof in front of him with the theorem to be proved. This comparison immediately leads to a classification of each last line of proof into one of four categories: additional 1's need to be added to master the theorem (R1); additional 0's need to be added to master the theorem (R2); a 1 must be deleted to continue to master the theorem (R3); or a 0 must be deleted in order to master the theorem (R4). The rule in terms of whether the response should be made is shown in parentheses. When the subject is completely conditioned to all four stimulus discriminations, he will make a correct response corresponding to the application of a rule that will produce a part of a proof of minimal length. For each of the four discriminations with respect to which he is not yet conditioned, there is a guessing probability p_i, $i = 1, 2, 3,$ or 4, that he will guess the correct rule and thus the probability $1 - p_i$ that he will guess incorrectly. In the analysis of data it was assumed that four independent one-element models were applied, one for each stimulus discrimination. (It is a minor but not serious complication to take account of two possible responses, both correct, i.e., leading to a minimal proof; e.g., in the proof of 1111 we may apply R1 twice and then R3, or R1, R3, and then R1 again.)

The pilot study was conducted with a group of first-grade children from an elementary school near Stanford University. There were 18 subjects in all divided into 2 groups of 9 each. One group received the procedure just described, including a correction procedure in terms of which a correct response was always shown at the end of the trial. The other group used a discovery method of sorts and was not given a correction procedure on each trial but, at the end of each proof, the subjects were shown a minimal proof or, in the event the subject constructed a minimal proof, told that the proof constructed was correct.

The following criterion rule was used: A subject, according to the criterion, had learned how to give minimal proofs in the system when 4 correct theorems were proved in succession, provided the subject had proved at least 10 theorems. All subjects were given a maximum of 17 theorems to prove, and all subjects, except for 2 in the discovery group, satisfied this criterion by the time the seventeenth theorem was reached. The 17 theorems were selected according to some relatively definite criteria of structural simplicity from the set of theorems of which the length was greater than 1 and less than 7.

In Table 6, the mean proportion of errors prior to the last error, in blocks of 12 trials for each group and for the 2 groups combined, are summarized. A trial in this instance is defined as a step, or line, in the proof.

More than 60 trials were necessary in order to prove the 17 theorems, but because very few subjects needed the entire 17 theorems to reach criterion, the mean learning curves were terminated at Trial 60. From this table, it seems that the correction group did better than the discovery group, but I do not think the number of subjects or the total number of trials was adequate to draw any serious conclusions about comparison of the two methods. It is interesting to note that the discovery group had a

TABLE 6

OBSERVED PROPORTION OF ERRORS PRIOR TO LAST ERROR FOR THE CORRECTION, DISCOVERY, AND COMBINED GROUPS (BLOCKS OF 12 TRIALS)

	Block				
Group	1	2	3	4	5
Correction..........	.28	.23	.15	.00	.10
Discovery..........	.23	.20	.40	.30	.33
Combined..........	.25	.21	.30	.18	.24

much more stationary mean learning curve than did the correction group, and in that sense satisfied the one-element model. Of course, these curves are obtained by summing over errors on all four rules. It is very possible that with a larger set of data, for which it would be feasible to separate out the individual rules as the application of the one-element model described above would require, the correction group also would have stationary mean learning curves for data prior to the last error on the basis of the individual rules.

REFERENCES

Bower, G. H. Application of a model to paired-associate learning. *Psychometrika,* 1961, 25, 255–280.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. *A study of thinking.* New York: John Wiley & Sons, 1956.

Burke, C. J., Estes, W. K., & Hellyer, S. Rate of verbal conditioning in relation to stimulus variability. *J. exp. Psychol.,* 1954, 48, 153–161.

Coombs, C. H. Psychological scaling without a unit of measurement. *Psychol. Rev.,* 1950, 57, 145–158.

Estes, W. K. The statistical approach to learning theory. In S. Koch (Ed.), *Psychology: a study of a science.* Vol. 2. New York: McGraw-Hill, 1960. Pp. 380–491.

Hill, S. A study of the logical abilities of children. Unpublished Ph.D. dissertation, Stanford University, 1961.

Hull, C. L., & Spence, K. W. Correction vs. non-correction method of trial-and-error learning in rats. *J. comp. Psychol.,* 1938, 25, 127–145.

Murphy, J. V., & Miller, R. E. The effect of spatial contiguity of cue and reward in the object-quality learning of Rhesus monkeys. *J. comp. physiol. Psychol.,* 1955, 48, 221–229.

Murphy, J. V., & Miller, R. E. Spatial contiguity of cue, reward and response in discrimination learning by children. *J. exp. Psychol.,* 1959, 58, 485–489.

Restle, F. Sources of difficulty in learning paired-associates. *In* R. C. Atkinson (Ed.), *Studies in Mathematical Psychology.* Stanford, Calif.: Stanford University Press, 1964. Pp. 116–172.

Stoll, E. Geometric concept formation in kindergarten children. Unpublished Ph.D. dissertation, Stanford University, 1962.

Suppes, P. Towards a behavioral foundation of mathematical proofs. Technical Report No. 44, Psychology Series, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1962.

Suppes, P., & Atkinson, R. C. *Markov learning models for multiperson interactions.* Stanford, Calif.: Stanford University Press, 1960.

Suppes, P., & Ginsberg, R. A fundamental property of all-or-none models, binomial distribution of responses prior to conditioning, with application to concept formation in children. *Psychol. Rev.*, 1963, 70, 139–161.

Suppes, P., & Ginsberg, R. Application of a stimulus sampling model to children's concept formation with and without an overt correction response. *J. exp. Psychol.*, 1962, 63, 330–336. (a)

Suppes, P., & Ginsberg, R. Experimental studies of mathematical concept formation in young children. *Sci. Educ.*, 1962, 46, 230–240. (b)

Suppes, P., & Hill, S. *First Course in Mathematical Logic.* New York: Blaisdell Publishing Co., 1964.